Attorney Docket No. 826.1894

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:

Masao OOTA

Application No.: (Unassigned) Group Art Unit:

Filed: (Concurrently) Examiner:

For: METHOD FOR STORING DATA USING GLOBALLY DISTRIBUTED STORAGE SYSTEM, AND PROGRAM AND STORAGE MEDIUM FOR ALLOWING COMPUTER TO REALIZE THE METHOD, AND CONTROL APPARATUS IN GLOBALLY DISTRIBUTED STORAGE SYSTEM

SUBMISSION OF CERTIFIED COPY OF PRIOR FOREIGN APPLICATION IN ACCORDANCE WITH THE REQUIREMENTS OF 37 C.F.R. § 1.55

Commissioner for Patents
PO Box 1450
Alexandria, VA 22313-1450

Sir:

In accordance with the provisions of 37 C.F.R. § 1.55, the applicant(s) submit(s) herewith a certified copy of the following foreign application:

Japanese Patent Application No(s). 2002-286528

Filed: September 30, 2002

It is respectfully requested that the applicant(s) be given the benefit of the foreign filing date(s) as evidenced by the certified papers attached hereto, in accordance with the requirements of 35 U.S.C. § 119.

Respectfully submitted,

STAAS & HALSEY LLP

Date: 9/22/03

By: Richard A. Gollhofer
Richard A. Gollhofer
Registration No. 31,106

1201 New York Ave, N.W., Suite 700
Washington, D.C. 20005
Telephone: (202) 434-1500
Facsimile: (202) 434-1501

日　本　国　特　許　庁
JAPAN　PATENT　OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application:　２００２年　９月３０日

出願番号
Application Number:　特願２００２－２８６５２８

[ST.10/C]:　[ＪＰ２００２－２８６５２８]

出願人
Applicant(s):　富士通株式会社

２００３年　１月２４日

特許庁長官
Commissioner,
Japan Patent Office　太田信一郎

出証番号　出証特２００３－３００１４４２

JAPAN PATENT OFFICE

This is to certify that the annexed is a true copy of the following application as filed with this Office.

Date of Application: September 30, 2002

Application Number: Patent Application No. 2002-286528
[ST.10/C]: [JP2002-286528]

Applicant(s): FUJITSU LIMITED

January 24, 2003

Commissioner,
Japan Patent Office Shinichiro OTA

Certificate No. P2003-3001442

【書類名】　特許願

【整理番号】　0252064

【提出日】　平成14年 9月30日

【あて先】　特許庁長官殿

【国際特許分類】　G06F　3/06

【発明の名称】　広域分散ストレージシステムを利用したデータ格納方法、その方法をコンピュータに実現させるプログラム、記録媒体、及び広域分散ストレージシステムにおける制御装置

【請求項の数】　5

【発明者】

【住所又は居所】　神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

【氏名】　太田　昌男

【特許出願人】

【識別番号】　000005223

【氏名又は名称】　富士通株式会社

【代理人】

【識別番号】　100074099

【住所又は居所】　東京都千代田区二番町８番地２０　二番町ビル３Ｆ

【弁理士】

【氏名又は名称】　大菅　義之

【電話番号】　03-3238-0031

【選任した代理人】

【識別番号】　100067987

【住所又は居所】　神奈川県横浜市鶴見区北寺尾７－２５－２８－５０３

【弁理士】

【氏名又は名称】　久木元　彰

【電話番号】　045-573-3683

【手数料の表示】

【予納台帳番号】　012542

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9705047

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　広域分散ストレージシステムを利用したデータ格納方法、その方法をコンピュータに実現させるプログラム、記録媒体、及び広域分散ストレージシステムにおける制御装置

【特許請求の範囲】

【請求項１】　コンピュータが、データを冗長化して複数のボリウムに分割し、各ボリウムを、ネットワークを介して分散配置された複数のストレージに分散して格納するデータ格納方法であって、

帯域幅、通信コスト及び書き込みを依頼するノードからストレージまでの経路の物理的距離に基づいて、前記分散配置された各ストレージについて利用対象としての望ましさを示す評価値を算出し、

前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択する、

ことを特徴とするデータ格納方法。

【請求項２】　前記データを読み込む際には、各ストレージについて、前記帯域幅及び前記コストに基づいてレスポンスの良さを示す利用優先度を算出し、

前記利用優先度に基づいて冗長化部分を含まないボリウムとして、前記複数のボリウムのうちいずれのボリウムを各ストレージから読み出すか決定する、

ことを更に含むことを特徴とする請求項１に記載のデータ格納方法。

【請求項３】　前記ストレージセットとして選択されなかったストレージに、前記複数のボリウムうちの第１のボリウムの複製を格納する、

ことを更に含むことを特徴とする請求項１に記載のデータ格納方法。

【請求項４】　ネットワークを介して分散配置されたストレージを備えるシステムにデータを冗長化して複数のボリウムに分割し、各ボリウムを複数のストレージに分散して格納する制御をコンピュータに行われるコンピュータ・プログラムであって、

帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離に基づいて、前記分散配置された各ストレージについて利用対象としての望ましさを示す評価値を算出し、

前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択する、

ことを含む制御を前記コンピュータに行わせることを特徴とするコンピュータ・プログラム。

【請求項５】　ネットワークを介して分散配置されたストレージを備えるシステムにデータを冗長化して複数のボリウムに分割し、各ボリウムを複数のストレージに分散して格納する制御を行う制御装置であって、

帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離に基づいて、前記分散配置された各ストレージについて利用対象としての望ましさを示す評価値を算出する経路管理手段と、

前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択するストレージセット管理手段と、

を備えることを特徴とする制御装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、システム内の格納装置を多重化することにより、データの冗長性や格納装置の性能の改善を図る技術、つまりＲＡＩＤ（Redundant Array of Inexpensive Disks）に係わる技術に関する。

【０００２】

【従来の技術】

従来、ＲＡＩＤを用いて、１つのデータを複数に分割し、分割されたデータを複数のストレージへ分散して格納することにより、システムの耐障害性（フォールトトレランス性）を向上させることが行われている。ＲＡＩＤには、レベル０からレベル６までの７レベルがあり、さらに、複数のＲＡＩＤのレベルを組み合わせたレベルや共通化されていない独自レベルもある。このうち、レベル５は、データを複数に分割し、それらの分割されたデータの各々にパリティデータを付加し、それぞれの分割されたデータを複数のストレージに分散させて格納する。レベル５は、１台のネットワーク端末もしくは処理サーバと、ファイルサーバの

ように、比較的密な関係を持った装置間で採用されることが好ましい。

【０００３】

図３３に、ＲＡＩＤを採用したシステムの構成図を示す。まず、図３３において、ルータＲ（中継器）を用いて、ストレージサービスセンタＳＣ、バックアップセンタＢＣ、ミラーセンタＭＣ及び利用者（ネットワーク端末あるいは処理サーバ）が接続され、これによって広域分散ストレージシステムが形成されている。

【０００４】

以下、利用者が本社と支社とに分かれており、本社の利用者ＵＨがデータを広域分散ストレージシステムにデータを格納し、そのデータを支社の利用者ＵＢが利用すると仮定して、この広域分散ストレージシステムにおいて行われる処理について説明する。

【０００５】

まず、本社の利用者ＵＨは、格納したいデータをストレージサービスセンタＳＣ内のストレージに格納させる。ストレージサービスセンタＳＣは、データを複製し、複製されたデータをバックアップセンタＢＣ内のストレージに格納させる。なお、災害等によってストレージサービスセンタＳＣとバックアップセンタＢＣの双方が損害を受ける可能性を低くするために、ストレージサービスセンタＳＣとバックアップセンタＢＣとの物理的な距離は遠隔であることが望ましい。

【０００６】

さらに、支社の利用者ＵＢがデータをストレージから読み出す際のレスポンスを改善するために、ストレージサービスセンタＳＣはデータを複製し、複製されたデータを、支社に最寄りの接続地点となっているミラーセンタＭＣ内のストレージに格納させる、又は、支社の利用者ＵＢからストレージサービスセンタＳＣまでの回線に割り当てる帯域を広くする。なお、バックアップセンタＢＣがミラーセンタＭＣを兼ねることとしてもよい。

【０００７】

また、ＲＡＩＤに関する技術として、データをセグメントに分割し、そのセグメントごとに複数のストレージにランダムに分散格納する、つまり、ストライピ

ング先となるストレージをランダムに決定する第１の発明がある。第１の発明により、一次ストレージが故障した場合その負荷全体が二次バックアップストレージにかかるという問題、及びコンボイ効果の確度が高くなるという問題を解決することが可能となる（例えば、特許文献１参照）。

【０００８】

また、さらなるＲＡＩＤに関する技術として、ストレージに格納されたデータをミラーリングする際に、そのデータを複数に分割し、分割されたデータを複数のストレージに分散して格納する第２の発明がある。第２の発明により、元となったストレージに障害が生じた場合でも、複数のストレージに分散格納されていたデータを読み出して、これらのデータを用いて、元となったストレージに格納されていたデータを復元することが可能となる（例えば、特許文献２参照）。

【０００９】

また、さらなるＲＡＩＤに関する技術として、バッファに格納されたデータをストレージに書き出す際に、送出先となるストレージが同一であるデータが複数存在する場合、それらのデータを１つの複合パケットにまとめて送出する第３の発明がある。第３の発明により、ＲＡＩＤのＩ／Ｏスループット性能を向上させることが可能となる（例えば、特許文献３参照）。

【００１０】

【特許文献１】

特表２００２－５００３９３号公報（段落０００５から段落０００７、図１）

【００１１】

【特許文献２】

特開平９－１７１４７９号公報（段落００１８から段落００２０、段落０３１から段落００３４、図１）

【００１２】

【特許文献３】

特開平１０－３３３８３６号公報（段落００２４から段落００２７、図３）

【００１３】

【発明が解決しようとする課題】

しかし、上記の図３３に示した従来に係わる広域ネットワークシステムには、以下に挙げる問題があった。

【００１４】

１）　バックアップセンタ及び・又はミラーセンタに、ストレージサービスセンタＳＣと同容量のストレージを備えることが必要であるため、システムが高コストとなる。

【００１５】

２）　バックアップ又はミラーリングを行う場合、そのために回線を利用することとなり効率的ではない。

３）　利用者が利用するネットワーク端末或いは処理サーバが、マルチホーミングされている場合であっても、それによって利用可能となっている複数の回線を効率的に使用できない。

【００１６】

４）　各センタＳＣ、ＢＣ又はＭＣ内のストレージ等が盗難された場合、ストレージに格納されていたデータが損害を受けるため、セキュリティがよくない。

また、上記に記載の第１の発明から第３の発明においても、上述の１）から４）の問題が解決されていない。

【００１７】

以上の問題に鑑み、データの冗長化に要するストレージ容量を低減しつつも、データのセキュリティを向上させ、且つ、回線を効率的に利用することが可能なＲＡＩＤを提供することが、本発明が解決しようとする課題である。

【００１８】

【課題を解決するための手段】

上記問題を解決するために、本発明の１態様によれば、コンピュータが、データを冗長化して複数のボリウムに分割し、各ボリウムを、ネットワークを介して分散配置された複数のストレージに分散して格納するデータ格納方法において、帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距

離に基づいて、前記分散配置された各ストレージについて利用対象としての望ましさを示す評価値を算出し、前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択することを含む。

【００１９】

データを複数のボリウムに分割して複数のストレージに分散して格納することによりデータのセキュリティを向上させつつ、さらに、帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離を考慮してそのノードから見て最適なストレージを選択することにより、回線効率と災害時のデータの安全性も向上させることが可能となる。

【００２０】

上記方法における、前記評価値の算出の際に、更に、前記書き込みを依頼するノードから各ストレージまでのホップ数を考慮する事としても良い。ホップ数がが高い場合は、回線効率が低下するからである。

【００２１】

また、上記方法において、前記システムの利用者に対して、前記ストレージセットを仮想的な１つのストレージとして提供することを更に含むこととしてもよい。これにより、データの分散格納によって、利用者にとって操作が複雑化することを避けることが可能となる。

【００２２】

また、上記方法において、前記データを前記ストレージセットから読み込む際には、前記ストレージセットに書きこまれた前記複数のボリウムのうち冗長化部分を含まないボリウムを各ストレージから読み出し、前記読み出されたボリウムを用いて前記データを復元する、ことを更に含むこととしてもよい。これにより、使用する回線帯域を抑制することが可能となる。

【００２３】

また、上記方法において、前記データを読み込む際には、前記帯域幅及び前記コストに基づいてレスポンスの良さを示す利用優先度を算出し、前記利用優先度に基づいて、冗長化部分を含まないボリウムとして、前記複数のボリウムのうち

いずれのボリウムを各ストレージから読み出すか決定することを更に含むこととしてもよい。例えば、データを３データ＋１パリティで冗長化して４つのボリウムに分割している場合、冗長化部分を含まないボリウムとして、３つのボリウムを任意に選択することができる。この選択の際に、帯域幅及びコストを考慮する事により、回線の利用効率を向上させることが可能となる。

【００２４】

また、上記方法において、前記ストレージセットとして選択されなかったストレージに、前記複数のボリウムうちの第１のボリウムの複製を格納することを更に含むこととしてもよい。この複製はバックアップとして用いることができる。従来、バックアップとしてデータの複製を備えていたため、バックアップを備えるためには元データの２倍の容量が必要であった。しかし、この態様によれば、バックアップのために必要なストレージ容量は、たかだか１ボリウム分の容量となる。これにより、ストレージの使用効率を向上させることが可能となる。

【００２５】

また、上記方法において、前記第１のボリウムの複製を作成する際に、前記評価値に基づいて、前記第１のボリウムを格納するストレージから前記第１のボリウムを複写するのか、前記複数のボリウムのうちの前記第１のボリウム以外のボリウムから冗長を利用して前記第１のボリウムを再現するのか、２つの作成方法のうちのいずれかを選択することを更に含むこととしてもよい。ここで、この選択において前記評価値を考慮することとしてもよい。

【００２６】

また、上記方法において、同一のボリウムを格納するべき複数のストレージに対して、マルチキャストでボリウムを書き込むこととしてもよい。これにより、同じ内容をもつパケットを複数回にわたり送信することを回避する。

【００２７】

また、上記方法において、前記第１のボリウムの複製をストレージに書き込む際に、多数回に分けて書き込み処理を行うこととしてもよい。多数回に分けることにより、一度につき回線にかける負荷を軽減する事が可能となる。

【００２８】

また、上記方法において、前記ストレージセットのうちの第１のストレージに障害が発生した場合、前記ストレージセットのうちの他のストレージへの書き込みを制限することを更に含むこととしてもよい。例えば、第１のストレージの復旧前に他のストレージ内のボリウムが更新されてしまう事がありうるが、これにより、障害が発生したストレージの復旧後、システム内に異なるバージョンのボリウムが並存する事を防止する。

【００２９】

また、上記方法において、前記ストレージセットのうち第３のストレージに障害が発生した場合、前記評価値に基づいて、前記ストレージセットとして選択されているストレージ以外の第４のストレージを、前記第３のストレージの代わりに選択することとしてもよい。これにより、障害が発生したストレージの代わりとして最適なストレージを選択することが可能となる。

【００３０】

また、上記方法において、前記ストレージセットの選択後、一定のタイミングで、各ノードにおけるストレージセットを再選択し、再選択の結果、どのノードからも利用されていないボリウムがあった場合、該ボリウムをストレージから削除することを更に含むこととしてもよい。ここで、一定のタイミングとは、前回の選択から一定期間後、又はボリウムの状態が変更される毎である。システムの使用状況が変化するタイミングにおいて、ボリウムの使用状況に基づいて不要なボリウムを削除する事により、ストレージの利用効率を向上させることが可能となる。

【００３１】

また、上記方法において、前記データを読み込んだ後に、前記データを一定期間、任意の１つのストレージ内に一時格納し、前記一定期間内にデータの読み出しを行う際には、一時格納されたデータを前記１つのストレージから読み出すことを更に含むこととしてもよい。このようなキャッシュ機能を備える事により、データの読み出しレスポンスを向上させることが可能となる。

【００３２】

また、上記方法において、一定期間内に書き込み要求されたデータを一時記憶

領域に保持し、前記一定期間経過後に一時格納領域からデータを取出し、データを複数のボリウムに分割し、該複数のボリウムを前記ストレージセットに書き込むことを更に含むこととしてもよい。これにより、書き込み要求を出すノードから他のストレージへボリウムを転送する回数を低減する事が可能となるため、トラフィックの効率を上げることが可能となる。

【００３３】

また、上記方法において、前記複数のストレージに前記複数のボリウムを書き込む際に、前記書き込みを依頼するノードは、書き込みが終了するまで前記複数のストレージへの書き込み処理を禁止することを更に含むこととしてもよい。ここで、同一のボリウムを格納するべき複数のストレージがある場合、それらのストレージの中から、１つのストレージを代表ストレージとして決定し、前記複数のストレージへの書き込み処理の禁止において、前記代表ストレージへの書き込み処理の禁止は、前記書き込みを依頼するノードによって行われ、前記代表ストレージ以外のストレージへの書き込み処理の禁止は、前記代表ストレージによって行われることとしてもよい。また、前記代表ストレージは、原本となるボリウムを格納するべきストレージであることとしてもよい。

【００３４】

また、上記方法に含まれる手順を含む制御をコンピュータに行わせるコンピュータ・プログラムも、コンピュータによって上記コンピュータ・プログラムを実行させる事によって、上記方法と同様の作用・効果が得られるため、上記課題を解決することが可能である。

【００３５】

また、上記コンピュータ・プログラムを記録したコンピュータ読み取り可能な記録媒体から、そのプログラムをコンピュータに読み出させて実行させることによっても、上記課題を解決することができる。

【００３６】

また、上記データ格納方法において行われる手順と同様の処理を行う、ネットワークを介して分散配置されたストレージを備えるシステムにデータを分散格納させる制御を行う制御装置によっても、上記データ格納方法と同様の作用・効果

が得られるため、上記課題を解決することが可能である。

【００３７】

【発明の実施の形態】

以下、本発明の実施の形態について図面を用いて説明する。なお、同じ装置等には同じ参照番号をつけ、説明を省略する。なお、以下の説明において、「ノードに備えられるストレージ」をいう際に「ノード」という場合がある。これは、文が長くなるために意味が分かりにくくなる事を防ぐためである。例えば、「ノードにボリウムを格納する」という表現は、「ノードに備えられるストレージにボリウムを格納する」ということを意味する。

【００３８】

本発明は、データにパリティを付加し、これらを複数のストレージに分散格納する技術、例えばＲＡＩＤ５等の技術を前提とする。図１に、本発明の各実施形態に係わる広域分散ストレージシステムの構成を示す。図１に示すように、広域分散ストレージシステムにおいて、ネットワークを介して複数のノードが接続されている。ノード間で通信されるデータは、ルータＲで中継される。各ノードは、ストレージＳ及び制御装置Ｃを備える。

【００３９】

利用者本社や利用者支店等に備えられた端末の利用者は、広域分散ストレージシステムにアクセスし、ストレージＳにデータを格納させたり、ストレージＳからデータを読み出したり等を行う。

【００４０】

制御装置Ｃは、端末からデータをストレージに格納するよう指示された場合、格納すべきデータのデータブロック単位（読み出し／書き込みの単位）にＥＣＣ(Error Check and Correct)／パリティを付し、複数のストレージＳにデータを分散させて格納させる。以下、分割されてパリティを付加されたデータをボリウムという。

【００４１】

端末からストレージ内に格納されたデータを読み出すよう指示された際には、制御装置Ｃは、複数のストレージＳから分散されて格納されているデータ、つま

りボリウムを読み出して、データを復元して端末に送信する。

【００４２】

データの格納及び読み出しの際、データの分散格納及び復元を制御装置Ｃが行うため、端末の利用者は、データが分散されていることを意識することなく、１つの仮想ディスクにデータを格納し、その仮想ディスクからデータを読み出す際と同様に、データを分散格納したり復元したりすることができる。

【００４３】

また、ボリウムを読み出してデータを復元する際に、制御装置Ｃは、そのデータを構成する全てのボリウムを複数のストレージから読み出してそのデータを復元する。或いは、制御装置Ｃは、冗長化された分のボリウムを除いたボリウムを複数のストレージＳから読み出してそのデータを復元する事としても良い。この場合、ネットワークにかかる負荷を低減する事ができる。より具体的には、２データ＋１パリティで冗長化されて分割されたデータを復元する場合、制御装置Ｃは、３つのボリウムのうち２つのボリウムを読み出してデータを復元する。

【００４４】

図２に、制御装置Ｃの構成を示す。図２に示すように、制御装置Ｃは、ユーザインタフェース（以下、ユーザＩＦ）（受信側）１、ユーザＩＦ（送信側）２、データ変換部３、パケット生成部４、制御部５、データ組立部６、パケット解析部７、ストレージインタフェース（以下、ストレージＩＦ）８、ネットワークインタフェース（以下、ネットワークＩＦ）（送信側）９及びネットワークＩＦ（受信側）１０を備える。

【００４５】

ユーザＩＦ（受信側）１は、利用者からストレージＳへアクセスするパケットを受信し、制御情報を制御部５に、データをデータ変換部３へ振り分ける。

データ変換部３は、データをデータブロックに分割し、各ブロックにパリティを付加する。

【００４６】

パケット生成部４は、広域ネットワークに送信するために、ブロック単位に分割されたデータ又は制御情報をパケット化する。

ネットワークＩＦ（送信側）９は、パケット生成部４によって生成されたパケットをネットワークに送信する。

【００４７】

ネットワークＩＦ（受信側）１０は、広域ネットワークからデータ又は制御情報を受信する。

パケット解析部７は、ネットワークＩＦ（受信側）１０から出力されたパケットを解析し、ストレージＳからのデータ読み出し又はストレージＳへのデータ書き込み処理を行う。

【００４８】

データ組立部６は、ストレージＳから読み出された信号を組み立てて、利用者からのデータアクセス指示に対して、制御情報を含む適正なパケットを生成する。

【００４９】

制御部５は、利用者からのアクセスに基づいて、ストレージＳ及びデータの管理及び送受信パケットを処理する。

ユーザＩＦ（送信側）２は、利用者に、データ組立部６によって組み立てられたパケットを送信する。

【００５０】

次に、図３に、制御装置Ｃの詳細構成図を示す。以下、図３に示す詳細構成図にそって、データ変換部３、パケット生成部４、制御部５及びデータ組立部６の動作について詳しく説明する。

【００５１】

上記データ変換部３は、パケット解析部３０１、データ分割部３０２及びパリティ計算部３０３を備える。パケット解析部３０１は、受信されたパケットを解析し、そのパケットからデータを取得する。データ分割部３０２は、データをデータブロック単位に分割する、パリティ計算部３０３は、パリティを計算し、データブロックに付加する。

【００５２】

上記パケット生成部４は、データ管理情報付加部４０１、制御／経路情報負荷

部４０２、データ転送部４０３及び転送パケット構築部４０４を備える。データ管理情報付加部４０１は、データブロックに制御部５からの出力されたデータ管理情報を付加する。なお、データ管理情報は、ストレージセット構成情報（後述）等であり、そのパケットに基づいて行われる処理の内容によって異なる。各処理において送信されるデータ管理情報については、後述する。

【００５３】

制御／経路情報付加部４０２は、データブロックに、制御情報や経路情報を付加する。なお、経路情報は、そのデータブロックの宛先となるノードまでの経路及びその経路の評価値を示す情報であり、制御部５によって生成される。制御情報は制御の内容、例えばデータの書き込みであるのか、読み込みであるのか、ストレージへの書き込みを制御するのか等を示す情報であり、制御部５によって生成される。

【００５４】

データ転送部４０３は、データ管理情報及び制御／経路情報が付加されたデータパケット又は、制御部５から出力された制御パケットを転送する。転送する際に、データ転送部４０３は、パケットの宛先となるノードのアドレス、例えばＩＰ（Internet Protocol）アドレス等をパケットに付加する。このアドレスは、制御部５から出力される。なお、制御／経路情報に基づいて、データが、そのノード内のストレージに書き込まれるべきデータ（ローカルデータ）であると判定された場合、データ転送部４０３は、そのデータをパケット解析部７に出力する。

【００５５】

転送パケット構築部４０４は、ストレージＩＦ８を介してストレージＳから読み出されたデータを他のノードの制御装置Ｃに転送するための転送パケットを構築し、データ転送部４０３に出力する。パケットを構築する際に、転送パケット構築部４０４は、上記のデータ管理情報付加部４０１及び制御経路情報付加部４０２と同様の処理を行う。

【００５６】

上記制御部５は、ストレージ制御部５０１、制御パケット生成部５０２、ネッ

トワーク制御部５０３、経路管理部５０４、ストレージセット管理部５０５、ローカルボリウム管理部５０６、経路評価テーブル５０７、ストレージ評価テーブル５０８、ストレージセット管理テーブル５０９、アクセス管理テーブル５１０及びローカルボリウム管理テーブル５１１を備える。ストレージ制御部５０１は、パケット解析部３０１から出力された制御情報に基づいて、ストレージＳへのデータ書き込み、読み出し及びロック等を制御する。また、ストレージ制御部５０１は、制御パケット生成部５０２、ネットワーク制御部５０３、経路管理部５０４、ストレージセット管理部５０５及びローカルボリウム管理部５０６の動作の連携も制御する。

【００５７】

制御パケット生成部５０２は、データ書き込み、データ読み込み、ストレージＳのロック等の制御の内容を示す制御パケットを生成する。この制御パケットは他のストレージへ送信される。ネットワーク制御部５０３は、経路管理部５０４からの出力に基づいて、パケットの宛先となるノードを示す経路情報及びそのノードのアドレス等を生成する。なお、ノードのアドレスは、不図示のアドレステーブルに登録されているとする。アドレステーブルについては自明であるため、ここでは説明しない。

【００５８】

経路管理部５０４は、経路評価テーブル５０７及びストレージ評価テーブル５０８に格納された情報に基づいて、ブロック単位に分割されたデータの宛先つまり、データの格納先又は転送先等を決定する。ストレージセット管理部５０５は、ストレージセット管理テーブル５０９を用いて、データを構成する複数のボリウムを管理する。また、ストレージセット管理部５０５は、各ノードのストレージＳに格納されたボリウムが更新される際に、アクセス管理テーブル５１０を用いて各ストレージへのアクセスを管理する。ローカルボリウム管理部５０６は、ローカルボリウム管理テーブル５１１を用いて、ローカルストレージの利用状況を管理する。各テーブルの構造について詳しくは後述する。

【００５９】

上記データ組立部６は、パケット構築部６０１、データ組立部６０２及びパリ

ティ計算部６０３を備える。パリティ計算部６０３は、パリティを計算する。データ組立部６は、パリティ及びボリウムがどのボリウムであるのかを示すボリウム番号に基づいて、ローカルストレージＳから読み出されたボリウムのデータ（ボリウムデータ）やその他のノードから受信したパケット内のボリウムデータを用いて、分割される前のデータを復元する。なお、ボリウム番号は、ボリウムデータに付されている。パケット構築部６０１は、データアクセス指示を出した利用者に、復元されたデータを送信するためにパケットを生成する。

【００６０】

次に、図４に、広域分散ストレージシステムの具体的な構成例を示す。以下、図４に示す具体的な構成を用いて各テーブルの構成や制御装置の動作等を説明するが、これは具体的に説明するために図４に示すような構成を仮定しているのであり、ストレージシステムの構成を限定する趣旨ではない。

【００６１】

図４に示すように、広域ネットワークを介して、ノードＡからノードＧから接続されている。各ノードには制御装置Ｃ及びストレージＳが備えられている。ノードＡとノードＢの間（以下、区間Ａ－Ｂ）、ノードＥとノードＦの間（以下、区間Ｅ－Ｆ）及び「ノードＦとノードＧの間（以下、区間Ｆ－Ｇ）の帯域幅は、１５０Ｍｂｐｓである。ノードＢとノードＣの間（以下、区間Ｂ－Ｃ）、ノードＣとノードＤの間（以下、区間Ｃ－Ｄ）及びノードＤとノードＥの間（以下、区間Ｄ－Ｅ）の帯域幅は、５０Ｍｐｂｓである。ノードＧとノードＡの間（以下、区間Ｇ－Ａ）の帯域幅は１Ｇｂｐｓである。ノードＢとノードＥの間（以下、区間Ｂ－Ｅ）の帯域幅は、６００Ｍｂｐｓである。

【００６２】

以下、図５から９を用いて、制御装置Ｃに備えられるテーブルの構造について説明する。まず、図５を用いて、経路評価テーブル５０７の構造について説明する。経路評価テーブル５０７は、広域分散ストレージシステムを構成する各ノードについての経路評価情報を格納する。経路評価テーブル５０７は、各ノード間を接続する経路の優位性を評価する際に参照される。図５に示すように、経路評価情報は、区間を識別する記号、その区間での帯域幅、その区間での通信コスト

、その区間の物理的距離（ディスタンス）、及びストレージの利用優先度等を項目として含む。また、経路評価情報は、さらに区間の利用優先度を更に含むこととしても良い。なお、ローカルノード、つまり、その経路評価テーブル５０７を備える制御装置Ｃが属するノードについては、ネットワークを介して通信する必要がないため、帯域幅、コスト及びディスタンスは空である。

【００６３】

帯域幅、通信コスト及び物理的距離は、広域分散ストレージシステムの構成に基づいて決定され、どのノードに備えられた制御装置であっても、基本的に同じ値となる。ストレージの利用優先度及び区間の利用優先度は、制御装置Ｃによって計算される値である。利用優先度は、帯域幅及びコストに加えて、災害等が発生した際のデータバックアップの安全性を評価するためにディスタンスに基づいて決定される。

【００６４】

ストレージ利用優先度の計算式は以下のとおりである。

ストレージ利用優先度

＝（帯域幅×Ａ）÷（コスト×Ｂ）＋（ディスタンス×Ｃ）

なお、Ａ、Ｂ及びＣは、重み付け定数である。以下の説明では、例として、Ａ、Ｂ及びＣをそれぞれ、２、１及び０．１と仮定する。各重み付け定数は、システムにおいて優先すべき性能、例えば通信速度を優先するのか、又はコストを優先するのか等を考慮して、変更することとしても良い。

【００６５】

図５は、図４にシステムのノードＡについての経路評価テーブル５０７を示している。以下、具体的に、区間Ａ－Ｂについてストレージの利用優先度を算出する。

【００６６】

区間Ａ－Ｂについてのストレージの利用優先度

＝（１５０×２）÷（１００×１）＋（８０×０．１）＝１１

従って、図５に示す経路評価テーブル５０７において、区間Ａ－Ｂについてのストレージの利用優先度として、「１１」が格納されている。

【００６７】

なお、区間の利用優先度は、帯域幅÷コストを正規化した値である。

次に、図６を用いて、ストレージ評価テーブル５０８の構造について説明する。ストレージ評価テーブル５０８は、広域分散ストレージシステムを構成する各ノードに備えられたストレージについてのストレージ評価情報を格納する。ストレージ評価テーブル５０８は、ストレージセットの作成及び追加等を行う際に、どのノードのストレージを使用すべきか決定する際に参照される。図６に示すように、ストレージ評価情報は、ノードを識別する記号、ローカルノードからそのノードへの経路、ストレージ評価値及びホップ数等を項目として含む。ここで、ローカルノードとは、そのストレージ評価テーブル５０８を備える制御装置Ｃのノードをいう。ホップ数とは、あるノードに到達するまでの経路に介在するノード数をいう。ストレージ評価値は、制御装置Ｃによって計算され、その計算式は以下の通りである。

【００６８】

ストレージ評価値＝

Σ｛（経路上のノードのストレージの利用優先度）×（重み付け定数）｝

／（その経路の最終ノードまでのホップ数）

ここで、重み付け定数をホップ数の逆数とすると、ストレージ評価値の計算式は以下のようになる。

【００６９】

ストレージ評価値

＝Σ｛（経路上のノードのストレージの利用優先度）÷（そのノードまでのホップ数）｝／（その経路の最終ノードまでのホップ数）

図６は、図４に示すシステム内のノードＡに備えられるストレージ評価テーブル５０８を例示している。以下、例として、具体的に、ノードＡから見たノードＢのストレージの評価値及びノードＣのストレージの評価値を算出する。

【００７０】

ノードＢのストレージ評価値

＝（区間Ａ－Ｂのストレージの利用優先度÷１）／１

＝１１

ノードＣのストレージ評価値

＝｛（区間Ａ－Ｂのストレージの利用優先度÷１）＋（区間Ｂ－Ｃのストレージの利用優先度÷２）｝／２

＝｛１１÷１＋１７÷２｝／２

＝９．７５

また、ストレージ評価情報は、さらに、経路評価値を項目として含むこととしても良い。経路評価値は、最終ノードに到達するまでに経由する区間の利用優先度の和をホップ数で割算する事により計算される。

【００７１】

次に、図７を用いて、ストレージセット管理テーブル５０９の構造について説明する。ストレージセット管理テーブル５０９は、それぞれのストレージセットに関する情報を管理するためのテーブルである。ストレージセット管理テーブル５０９は、ストレージセット構成情報を格納する。ストレージセット構成情報は、ストレージセットを識別するストレージセット番号と、その番号に対応するストレージセットに関する情報、つまりプロパティを含む。

【００７２】

ここで、ストレージセットとは、システム全体から見ると、データを分割されたことにより得られた複数のボリウムを分散格納するストレージをいう。しかし、後述のように、通常、各ノードにおいて、ボリウムを分散格納するストレージの全てを使用（書き込み、読み出し等行う）するのではなく、使用するストレージはこれらのうちの少なくとも一部である。各ノードで使用するストレージは、後述のプロパティに含まれる使用状況情報によって管理される。使用するストレージ以外は、バックアップ等の機能を果たす。従って、各ノードから見ると、ストレージセットとは、使用状況情報によって使用が許可されているストレージである。

【００７３】

ストレージセット番号は、ストレージセットを識別するために用いられるが、データを識別する情報としても利用される。例えば、システムの利用者は、読み

出したいデータを特定するためにストレージセット番号を用いる。これは、ストレージセットは、利用者に対して１つの仮想的なストレージとして提供されるからである。

【００７４】

プロパティには、広域分散ストレージシステム全体のプロパティと、各ノードのプロパティとがある。システム全体のプロパティは、データが分割されたノード数及びストレージの状態（良好であるか、異常があるのか等）を示す情報を含む。なお、データが分割されたノード数は、読み出しの場合と書き込みの場合のそれぞれについて格納される。なお、図７において、ストレージの状態として「Ｇ」が格納されている場合、状態が「良好」であり、「Ｒ」が格納されている場合、状態が「異常」である。

【００７５】

ノードのプロパティは、そのノードのストレージがストレージセット中のどのボリウムを格納するのかを示すボリウム番号、そのボリウムが原本であるのか複製であるのかを示すフラグ及び、ローカルノードから各ノードが読み出し可能なのか書き込み可能なのか等の使用状況を示す使用状況情報を含む。ストレージセット管理テーブル５０９内の情報は、各ノードに備えられた制御装置Ｃの間で交換される。なお、図７ではボリウムが原本であるのか複製であるのかを示すフラグが「Ｏ」である場合ボリウムは原本であり、「Ｃ」である場合ボリウムは複製である。なお、後述の逐次格納の場合、格納途中の不完全なボリウムデータがストレージＳに格納され得る。この不完全データを示すフラグを、「Ｏ」と「Ｃ」とは区別できるように、例えば「Ｑ」とすることとしてもよい。

【００７６】

例として、図７中のストレージセット番号が「００００００００１」であるストレージセット構造情報について説明する。このストレージセット構造情報の全体プロパティの読み出しノード数として「３」が格納されているため、データを復元するためには３つのボリウムが必要である。また、同様に、書き込みのノード数として「４」が格納されているため、データは４つのボリウムに分割されている。また、このストレージセット構造情報においてノードＡ、Ｂ、Ｃ、Ｅ及びＧ

についてのプロパティが書き込まれているため、これらのノードにボリウムが格納されている。例えば、ノードＡのプロパティによると、ノードＡにはボリウム番号が「１」であるボリウムが格納され、それは原本であり、読み出し（Read）及び書き込み（Write）が可能な状況であることがわかる。

【００７７】

また、ストレージセット構造情報中の使用状況情報は、そのストレージセット管理テーブル５０９を備えるノードにおいて通常使用されるストレージＳ、つまりそのノードから見たストレージセットを示す。本実施形態では、ローカルノードから見たストレージセットは、使用状況情報が「ＲＷ」つまり、読み出し及び書き込みが可能となっている複数のストレージであるとして仮定する。図７には、例としてノードＡに備えられるストレージセット管理テーブル５０９を示す。この図７によると、ストレージセット番号が「００００００００１」であるデータについては、ノードＡから見たストレージセットは、ノードＡ、Ｂ、及びＥに備えられたストレージＳである。

【００７８】

次に、図８を用いて、アクセス管理テーブル５１０のデータ構造について説明する。アクセス管理テーブル５１０は、アクセス単位ごとにアクセス管理情報を格納する。制御装置Ｃは、アクセス管理情報に基づいて、利用者からの各ノード内のストレージＳへのアクセスを制御する。同じストレージセットを共有する利用者からのアクセスは、同じ情報に基づいて制御される。アクセス管理情報は、ストレージセットアクセス番号及び、そのアクセス単位のプロパティを項目として含む。ストレージセットアクセス番号は、ストレージセットへの論理的なアクセス単位（論理ブロック）を示す番号であり、プロパティは、ストレージセットアクセス番号によって示される論理ブロックの状態、例えば、読み出し可能状態であるのか、書き込み可能状態であるのか、ロック状態であるのか、データが完全データであるのか。生成データであるのか等を示す。図８では、読み出し可能状態を「Ｒ」で例示し、書き込み可能状態を「Ｗ」で例示し、ロック状態を「Ｌ」で例示し、原本データを「Ｏ」で例示している。なお、ロック状態とは、書き込みが制限される状態をいい、例えばストレージ上のデータが更新される場合等

に、制御装置Ｃによって設定される（後述）。図８には、例として、ストレージセット番号が「０００００１０００１」であるストレージセットについてのアクセス管理テーブルが示されている。また、完全データとは、ストレージから読み出したボリウムから復元されるデータであり、生成データとは、一部のボリウムが足りない状態で冗長化を利用して生成されたボリウムをいう。

【００７９】

なお、アクセス管理情報は、さらに、ロックキーを項目として含むこととしてもよい。ロックキーは、ストレージセット番号によって識別されるデータの更新を要求した利用者を識別するための情報であり、更新要求を受信した制御装置Ｃによって生成される。

【００８０】

次に、図９を用いて、ローカルボリウム管理テーブル５１１のデータ構造について説明する。ローカルボリウム管理テーブル５１１は、そのテーブルを備える制御装置Ｃに接続されているストレージＳ、つまりローカルストレージに格納されているボリウムの利用状況を管理するためのテーブルである。ローカルボリウム管理テーブル５１１は、各ノードの制御装置Ｃに個別に存在する。

【００８１】

ローカルボリウム管理テーブル５１１は、ローカルストレージへのアクセス単位ごとにボリウム管理情報を格納する。図９に示すように、ボリウム管理情報は、ローカルストレージへのアクセス単位を示すストレージアクセス番号、そのストレージアクセス番号が示す論理ブロックの状態を示すプロパティ、ストレージセットを識別するストレージセット番号、そのストレージアクセス番号に対応するストレージセットアクセス番号を項目として含む。

【００８２】

以下、広域分散ストレージシステムにおける動作について説明する。以下の説明において、データは、２データ＋１パリティの冗長構成をとって３つに分割されて、広域分散ストレージシステムに分散格納されると仮定する。しかし、これは説明を分かり易く、且つ、具体的にするためであり、データの冗長構成を限定する趣旨ではない。

【００８３】

広域分散システムの利用者は、通常、ネットワーク的に最も近くに位置するノードを介して、広域分散システムにアクセスする。

以下、図１０を用いて、利用者ＡがノードＡを介して広域分散システムにアクセスする場合のデータの流れについて説明する。図１０において、実線の矢印は利用者が感じるデータの流れを示し、破線の矢印は、実際のデータの流れを示す。利用者Ａが、ノードＡを介して広域分散システムにアクセスし、データの格納指示を出したと仮定する。

【００８４】

利用者Ａから見た場合、ノードＡに備えられた１つの仮想ディスクにデータが格納されたように感じられる。しかし、実際は、ノードＡの制御装置Ｃは、データにパリティを付して複数のボリウムに分割し、広域分散システムを構成する複数のノードに備えられたストレージＳに、ボリウムを分散させて書き込む。図１０の場合、ノードＡの制御装置Ｃは、データを３つのボリウムに分割し、それぞれを、ノードＡのストレージＳ（Ａ）、ノードＢのストレージＳ（Ｂ）及びノードＧのストレージＳ（Ｇ）に書き込む。

【００８５】

以下、データ書き込みの際の制御装置Ｃの動作についてより詳しく説明する。

１）制御装置Ｃのパケット解析部３０１は、受信されたパケットを解析し、そのパケットから書き込み指示を示す制御情報とデータを取得する。

【００８６】

２）データ分割部３０２は、データをデータブロック単位に分割する。

３）パリティ計算部３０３は、パリティを計算し、データブロックに付加する。

【００８７】

４）制御部５の経路管理部５０４は、データブロックを３つのボリウムに割り振る事によってデータを３分割し、任意の方法で３つのノードをそれぞれのボリウムを分散格納するべきストレージセットとして決定する。この説明では、ノードＡ、ノードＢ及びノードＧがストレージセットとして決定される。

【００８８】

５）ストレージセット管理部５０５は、ストレージセットの決定結果に基づいて、ストレージセット構成情報を作成し、ストレージセット管理テーブル５０９に書き込む。

【００８９】

６）制御パケット生成部５０２は、データ書き込み制御を指示する制御パケットを生成する。ネットワーク制御部５０３は、経路管理部５０４からの出力に基づいて、パケットの宛先となるノードＡ、Ｂ及びＧを示す経路情報及びそのノードのアドレス等を生成する。

【００９０】

７）データ管理情報付加部４０１は、３つのボリウムのデータブロックにストレージセット構成情報を付加し、制御経路情報付加部４０２は、データブロックに、書き込み制御を指示する制御情報及び経路情報を付加する。データ転送部４０３は、データパケット又は、制御部５から出力された制御パケットを転送する。

【００９１】

なお、制御／経路情報に基づいて、データがローカルストレージ、つまり、そのノードＡ内のストレージＳ（Ａ）に書き込まれるべきデータ（ローカルデータ）であると判定される場合、データ転送部４０３は、そのデータをパケット解析部７に出力する。

【００９２】

８）パケット解析部７は、複数のボリウムのうちの１つをローカルストレージＳ（Ａ）へ書き込む制御を行う。書き込み後、ローカルボリウム管理部５０６は、ボリウム管理情報を生成し、ローカルボリウム管理テーブル５１１に格納する。なお、ボリウム管理情報に含まれる値の内、プロパティ、ストレージセット番号は、パケットから読み出すことにより取得される。

【００９３】

９）転送されたボリウムは、各転送先のノードのパケット解析部７によってそのノードのストレージＳに書き込まれる。格納先のノードのローカルボリウム管

理部５０６は、上記と同様にしてボリウム管理情報を生成し、ローカルボリウム管理テーブル５１１に格納する。

【００９４】

一方、利用者が分散格納されたデータを読み出す場合、利用者Ａから見た場合、ノードＡに備えられた１つの仮想ディスクからデータが読み出されたされたように感じられる。しかし、実際は、ノードＡの制御装置Ｃは、複数のノードから３つのボリウムを読み出して、データを復元する。以下、データを読み出す場合の制御装置Ｃの動作についてより詳しく説明する。

【００９５】

１）制御装置Ｃのパケット解析部３０１は、受信されたパケットを解析し、そのパケットから読み出し指示を示す制御情報を取出す。

２）ストレージセット管理部５０５は、ストレージセット管理テーブル５０９から読み出し指示がされたデータのストレージセット構成情報を取得し、これにより、利用者がアクセスしているノードから見てストレージセットとなっているノードのノード名を取得する。この場合は、ノードＡ、Ｂ及びＧである。

【００９６】

３）ノードＡのデータ組立部６０２は、ローカルストレージＳからボリウムを取得する。

４）ノードＢ及びＧに格納されている残りの２つのボリウムを取得するために、制御パケット生成部５０２は、データ読み出し制御を指示する制御パケットを生成する。ネットワーク制御部５０３は、パケットの宛先となるノードＢ及びＧを示す経路情報及びそのノードのアドレス等を生成する。

【００９７】

５）データ管理情報付加部４０１は、制御パケットにストレージセット構成情報を付加し、制御経路情報付加部４０２は、データブロックに、書き込み制御を指示する制御情報及び経路情報を付加する。データ転送部４０３は制御パケットを転送する。

【００９８】

６）ノードＢ及びノードＧのそれぞれにおいて、転送パケット構築部４０４は

、制御パケットに基づいてストレージＳからボリウムを読み出し読み出されたデータをノードＡの制御装置Ｃに転送するための転送パケットを構築し、データ転送部４０３に出力する。データ転送部４０３はパケットをノードＡに転送する。

【００９９】

７）ノードＡのパケット解析部７は、受信したパケットから、ノードＢのストレージ及びノードＧのストレージから読み出された各ボリウムを取得する。

８）パリティ計算部６０３は、パリティを計算し、データ組立部６０２は、パリティ及びボリウム番号に基づいて、３つのボリウムから分割される前のデータを組み立てる。パケット構築部６０１は、データの読み出し指示を出した利用者に、組み立てられたデータを送信するためにパケットを生成する。

【０１００】

このようにしてデータを複数のボリウムに分割し、ネットワークを介して分散して存在する複数のストレージに各ボリウムを格納させることにより、以下の効果が得られる。

【０１０１】

・１つストレージが盗難にあった場合でも、そのストレージに格納されている１つのボリウムだけでは分割前の元のデータを復元することができないため、データの安全性が高くなる。

【０１０２】

・各ノード宛てのパケットはデータの一部でしかないため、ネットワーク経路においてパケットキャプチャリングを行っても、分割前の元のデータを復元することができない。

【０１０３】

・ストレージが分散配置されるため、ネットワーク的に負荷分散を行う事ができる。このため、従来の技術と同一の速度でバックボーンが構成されている場合は、データアクセスにかかる時間を短縮する事ができる。また、従来の技術と同一のレスポンスを維持する場合は、バックボーンに必要な帯域幅を低減することができる。

【０１０４】

・分散配置と保管が同時に行われるため、バックアップセンタを設けるより、ストレージの使用効率がよい。

上記において、複数のボリウムに分散して格納されたデータを構成する全てのボリウムを複数のノードのストレージＳから読み出してデータを復元する際の処理について説明した。しかし、複数のボリウムのうち冗長化された分のボリウムが無くともデータを復元する事ができるため、冗長化された分のボリウムを除いたボリウムを複数のノードのストレージＳから読み出すこととしても良い。より具体的には、２データ＋１パリティで冗長化されて３つのボリウムに分割されたデータの場合、３つのボリウムのうち２つのボリウムがあればデータを復元できるため、３つのボリウムのうち２つのボリウムをストレージＳから読み出して、データを復元することとしてもよい。この場合、ネットワークにかかる負荷をさらに低減する事ができる。

【０１０５】

ここで、冗長化された分のボリウムを除いたボリウムを複数のノードのストレージＳから読み出す場合、読み出されるボリウムの組合せは幾通りも考えられる。以下、このような場合に、最適なボリウムの組合せを決定する方法について説明する。

【０１０６】

この場合、読み出すべきストレージを決定するために、制御装置Ｃ内の経路評価テーブル５０７に格納される経路評価情報に区間の利用優先度を更に含み、ストレージ評価テーブル５０８に格納されるストレージ評価情報に経路評価値をさらに含む。

【０１０７】

以下、図１１を用いて利用優先度及び評価値の計算処理の手順について説明する。利用優先度（区間の利用優先度とストレージの利用優先度）及び評価値（経路評価値とストレージ評価値）は、ネットワーク構成が変更されたり、回線が断絶したり、ノードが追加・削除されたりして経路評価テーブル５０７に格納される情報が変更された場合に、全てのノードについて計算される。

【０１０８】

図１１に示すように、まず、経路管理部５０４は、利用優先度及び評価値の計算対象として１つのノードを取出し、そのノードがローカルノード（自ノード）であるか否か判定する（Ｓ１１）。計算対象のノードがローカルノードで無い場合（Ｓ１１：Ｎｏ）、Ｓ１２に進み、計算対象のノードがローカルノードである場合（Ｓ１１：Ｙｅｓ）、Ｓ１６に進む。

【０１０９】

Ｓ１２において、経路管理部５０４は、計算対象ノードからそのノードに隣接する他のノードまでの各区間について、その区間の帯域幅、コスト、距離を経路評価テーブル５０７から取得する。さらに、経路管理部５０４は、区間の利用優先度及びストレージの利用優先度を計算し、その計算結果に基づいて経路評価テーブル５０７を更新する（Ｓ１３）。なお、区間の利用優先度及びストレージの利用優先度の計算方法については既に説明した。

【０１１０】

経路管理部５０４は、計算対象ノードから他のノードまでの各経路について、経路評価値及びストレージ評価値を計算し、その計算結果に基づいてストレージ評価テーブル５０８を更新する（Ｓ１４）。さらに、経路管理部５０４は、全てのノードについて利用優先度及び評価値を計算したか否か判定する（Ｓ１５）。全てのノードについて計算を行った場合（Ｓ１５：Ｙｅｓ）、処理を終了し、そうでない場合、Ｓ１１にもどる。

【０１１１】

Ｓ１６において、経路管理部５０４は、利用優先度及び評価値を最大値に設定し（Ｓ１６）、Ｓ１５に進む。このように、利用優先度及び評価値を最大値に設定することにより、ボリウムの書き込み又は読み出しにおいて、ローカルノードのストレージＳは最優先されることになる。

【０１１２】

図１２（ａ）に、区間の利用優先度を含む、ノードＡについての経路評価テーブル５０７の一例を、図１２（ｂ）に、経路評価値を含む、ノードＡについてのストレージ評価テーブル５０８の一例を示す。図１２に示すテーブルがノードＡについてのテーブルであることは、ノードＡが「ローカル」として示されている

ことから分かる。なお、図１２（ａ）に示す経路評価テーブル５０７において、区間Ｃ－Ｄについての区間の利用優先度を基準として、他の区間の利用優先度は正規化されている。

【０１１３】

以下、図１２を用いて、経路評価値の計算方法について具体的に説明する。図１２（ａ）に示すように、区間Ａ－Ｂ及び区間Ｂ－Ｃについての区間の利用優先度は、それぞれ３及び２である。この場合、経路Ａ－Ｂ－Ｃについての経路評価値は以下のようにして算出される。

【０１１４】

経路Ａ－Ｂ－Ｃについての経路評価値

＝｛（区間Ａ－Ｂの区間の利用優先度）＋（区間Ｂ－Ｃの区間の利用優先度）｝÷（ホップ数）

＝（３＋２）÷２

＝２．５

従って、図１２（ｂ）において、経路Ａ－Ｂ－Ｃについての経路評価値として「２．５」が格納されている。

【０１１５】

データを復元する際に、経路管理部５０４は、ストレージセット管理テーブル５０９から復元すべきデータのボリウムを格納するストレージＳを備えるノードのノード名を取得し、それらのノードのうち、ストレージ評価テーブル５０８内の経路評価値が大きいノードのストレージＳから優先してボリウムを読み出す事として決定する。読み出しボリウムの読み出しの際は、保管の安全性を考慮する必要は無いため、ディスタンスが考慮されていない経路評価値に基づいてボリウムを読み出すべきストレージを決定することは合理的である。

【０１１６】

以下、より具体的に、ノードＡ、Ｂ及びＧのストレージに３つのボリウムを分散格納した場合に、経路管理部５０４がボリウムを読み出すべきストレージを決定する方法について図１２（ｂ）を用いて説明する。ここで、ノードＡにアクセスした利用者に復元したデータを送信すると仮定する。

【０１１７】

図１２（ｂ）に示すように、ノードＡ、Ｂ及びＧについての経路評価値は、それぞれ「最大」、「３」及び「１０」である。３つのボリウムのうち２つのボリウムがあればデータを復元する事ができるので、この場合、ノードＡに備えられた制御装置Ｃ内の経路管理部５０４は、ノードＡのストレージ及びノードＧのストレージから１つずつボリウムを読み出すことを決定する。これにより、ネットワーク上の使用帯域を削減しつつ、良いレスポンスでボリウムをストレージから読み出して、データを復元することが可能となる。

【０１１８】

上記広域分散ストレージシステムにデータを分散させて格納する場合に、ノードの数がボリウムの数よりも多いことがよくある。この場合、いずれのノードのストレージＳにボリウムを格納するのか選択することが可能である。以下、最適なストレージセットを決定する方法について説明する。

【０１１９】

まず、基本的な考え方について説明する。

ボリウムをネットワーク上に分散されたストレージに格納する場合、帯域幅、コスト及びノード間のディスタンスを考慮することが望ましい。つまり、回線帯域が広く、コストが安いことに加えて、災害からの早期復旧のためにノード間の物理的な距離が離れていることが望ましい。近接したノードの場合、１つの災害のため同時に障害が発生する事がありうるからである。経路評価テーブル５０７に格納されるストレージ利用優先度及びストレージ評価テーブル５０８に格納されるストレージ評価値は、上記の考え方に基づいて、回線帯域が広い程大きな値となり、コストが安い程大きな値となり、ノード間の物理的な距離が離れている程大きな値となるように定義されている。なお、ストレージ利用優先度及びストレージ評価値の計算方法については既に説明した。

【０１２０】

以下、図１３を用いて、上記ストレージ評価値に基づいてボリウムを格納するストレージセットを決定する処理について説明する。この処理は利用単位ごとに行われる。なお、以下の説明において、ストレージ評価テーブル５０８に経路評

価値が項目として含まれていると仮定する。

【０１２１】

利用者から新規のデータの格納指示を受けた場合、ストレージセットを新規に決定することが必要となる。なお、ここでいうストレージセットとは、利用者がアクセスしているノード、つまりローカルノードから見たストレージセットである。ストレージセットとして決定されるべきノードの数は、通常、データを分割する事により得られたボリウムの数と同数である。

【０１２２】

ストレージセットを決定するために、まず、ローカルノードの経路管理部５０４は、ストレージセット番号を割り当て、ストレージセット構成情報をストレージセット管理テーブル５０９に格納する（Ｓ２１）。なお、この時点では、ストレージセット構成情報にはストレージセット番号が割り当てられているだけであり、中身は空である。

【０１２３】

つづいて、経路管理部５０４は、ストレージ評価テーブル５０８を参照し、まだストレージセットを構成するノードとして決定されていないノードの中から、最大のストレージ評価値と、その評価値を持つノードのノード名を取得する（Ｓ２２）。

【０１２４】

経路管理部５０４は、Ｓ２２において同一のストレージ評価値を持つ複数のノードを取得したか否か判定する（Ｓ２３）。同一のストレージ評価値を持つ複数のノードを取得した場合（Ｓ２３：Ｙｅｓ）、Ｓ２４に進み、そうでない場合（Ｓ２３：Ｎｏ）、Ｓ３０に進む。

【０１２５】

Ｓ２４において、経路管理部５０４は、更に、同一のストレージ評価値を持つノードの数は、不足しているノードの数以上であるか否か判定する。同一のストレージ評価値を持つノードの数が、不足しているノードの数以上である場合（Ｓ２４：Ｙｅｓ）、Ｓ２５に進み、そうでない場合（Ｓ２４：Ｎｏ）、Ｓ３１に進む。なお、不足しているノード数とは、ストレージセットを構成するノードとし

て決定されるべきノードの数から、ストレージセットを構成するノードとして決定されたノードの数を減算した後に残る数である。つまり、不足しているノード数とは、ストレージセットを構成するノードとして決定すべきノードの総数のうち、まだ決定されていないノードの数をいう。

【０１２６】

Ｓ２５において、経路管理部５０４は、Ｓ２２で取得された複数のノードについて、ローカルノードからそのノードに至るまでのホップ数が互いに同じであるか否か判定する。ホップ数が互いに同じである場合（Ｓ２５：Ｙｅｓ）、Ｓ２６に進み、そうでない場合（Ｓ２５：Ｎｏ）、Ｓ３２に進む。

【０１２７】

Ｓ２６において、経路管理部５０４は、ストレージセット管理テーブル５０９からＳ２２で取得された複数のノードの経路評価値を取得し、これらのノードの経路評価値が互いに同じであるか否か判定する。複数のノードの経路評価値が互いに同じである場合（Ｓ２６：Ｙｅｓ）、Ｓ２７に進み、そうでない場合（Ｓ２６：Ｎｏ）、Ｓ３３に進む。

【０１２８】

Ｓ２７において、経路管理部５０４は、Ｓ２２で取得した複数のノードから不足しているノード数と同数のノードを任意に選択し、各ノードのストレージに格納するボリウムを決定する。そして、経路管理部５０４は、ストレージセット構成情報中の各ノードに対応するフィールドに決定したボリウム番号を書き込む。その際に、経路管理部５０４は、ボリウムが原本であるか否かを示すフラグ（この場合は原本）及び使用状況情報（書き込み可能及び読み込み可能）も書き込む。これにより、Ｓ２２で取得されたノードはストレージセットを構成するノードとして決定され、ノードの状態は利用状態となる。

【０１２９】

続いて、経路管理部５０４は、ストレージセットを構成するために必要な数だけノードを決定したか否か判定する（Ｓ２８）。必要な数の分だけノードを決定した場合（Ｓ２８：Ｙｅｓ）、処理を終了し、そうでない場合（Ｓ２８：Ｎｏ）、Ｓ２２にもどる。

【０１３０】

Ｓ３０において、経路管理部５０４は、Ｓ２２で取得されたノードを、ストレージセットを構成するノードとして決定する。経路管理部５０４は、Ｓ２７と同様にして、そのノードに備えられたストレージに書き込むボリウムのボリウム番号を決定し、決定結果、ボリウムが原本であるか否かを示すフラグ及び使用状況情報をＳ２１で作成したストレージセット構成情報に書き込む。その後、Ｓ２８に進む。

【０１３１】

Ｓ３１において、経路管理部５０４は、Ｓ２２で取得された複数のノードを、ストレージセットを構成するノードとして決定する。経路管理部５０４は、Ｓ２７と同様にして、Ｓ２１で作成したストレージセット構成情報に決定結果、フラグ及び使用状況情報を書き込む。その後、Ｓ２８に進む。

【０１３２】

Ｓ３２において、経路管理部５０４は、Ｓ２２で取得された複数のノードのうち、ホップ数が少ない方のノードを、ストレージセットを構成するノードとして決定する。経路管理部５０４は、Ｓ２７と同様にして、Ｓ２１で作成したストレージセット構成情報に決定結果、フラグ及び使用状況情報を書き込む。その後、Ｓ２８に進む。

【０１３３】

Ｓ３３において、経路管理部５０４は、Ｓ２２で取得された複数のノードのうち、経路評価値が大きい方のノードを、ストレージセットを構成するノードとして決定する。経路管理部５０４は、Ｓ２７と同様にして、Ｓ２１で作成したストレージセット構成情報に決定結果、フラグ及び使用状況情報を書き込む。その後、Ｓ２８に進む。

【０１３４】

上記のようにして経路管理部５０４はストレージセットを構成するノードを決定し、ストレージセット構成情報が作成される。決定結果に基づいて、複数のボリウムは、ストレージセットとして決定されたノードに備えられたストレージに分散格納される。また、このストレージセット構成情報は、各ノードの制御装置

Ｃに送信され、ストレージセット管理テーブル５０９に格納される。なお、ストレージ評価テーブル５０８に経路評価値が含まれていない場合、上記処理において、Ｓ２６及びＳ３３は行われない。ストレージセットは、ネットワーク構成が変化した際等に、更新することが可能である。この場合、経路管理部５０４は、更新されるストレージセットについてのストレージセット構造情報の使用状況情報をクリアした後、Ｓ２２以降を行う。

【０１３５】

上記のように、制御装置Ｃは、データを複数のボリウムに分割し、帯域幅やコストだけでなく、ノード間の物理的距離に基づいて選択されたストレージにそれらのボリウムを格納する。これにより、災害により１つのボリウムを格納するストレージが破壊等された場合であっても、他のストレージに格納されたボリウムが無事であれば、データを復元する事が可能となる。従って、十分にノード間の物理的距離があれば、特にデータをバックアップするためのバックアップセンタを備える必要が無くなるという効果が得られる。

【０１３６】

以下、３データ＋１パリティで冗長化したデータを複数のストレージに分散格納する場合を例として、ストレージセットの決定方法について具体的に説明する。なお、利用者はノードＡにアクセスしていると仮定する。

【０１３７】

この場合、格納されるべきデータは、４つのボリウムに分割される。従って、ノードＡに備えられた制御装置Ｃ内の経路管理部５０４は、ストレージ評価テーブル５０８に格納されたストレージ評価値に基づいて、これらのボリウムを分割格納するべき４つのストレージを決定する。図１４（ａ）に、経路評価テーブル５０７の一例を、図１４（ｂ）に、図１４（ａ）に示す経路評価テーブル５０７中のデータに基づいて算出されたストレージ評価値を示す。

【０１３８】

図１４に即して説明すると、経路管理部５０４は、最も大きなストレージ評価値を持つノードのストレージから順に４つのストレージ、つまり、ノードＡ、Ｂ、Ｅ及びＧのストレージを、ボリウムを格納するべきストレージとして決定する

。経路管理部５０４は、決定結果に基づいてストレージセット構成情報を作成する。

【０１３９】

上記のように、冗長化され分割されたデータを構成する各ボリウムは、広域分散ストレージシステムに分散して格納される。さらに、各ボリウムの複製を作成し、ストレージに格納することとしても良いことはいうまでも無い。

【０１４０】

次に、ストレージセットを決定する際にローカルノードとなったノード以外のノードにアクセスする利用者が、そのストレージセットに格納されるデータを利用する場合について説明する。

【０１４１】

以下、データの格納を指示し、ストレージセットを決定する際にアクセスした利用者を利用者Ａ、その利用者ＡがアクセスするノードをノードＡ、そのストレージセットに格納されるデータを利用する新たな利用者を利用者Ｅ、その利用者ＥがアクセスするノードをノードＥと仮定して説明する。

【０１４２】

利用者Ｅは、ノードＥを介して、広域分散ストレージシステムからデータを取得することができるが、上記のストレージセットは、ノードＡから見て利用効率が良いように最適化されている。そこで、新たな利用者Ｅが利用するノードＥから見ても良好な利用効率が得られるように、ボリウムの複製を作成することが可能である。以下、この処理を利用者の追加処理という。

【０１４３】

図１５に、ストレージセットを利用する利用者を追加する処理の手順を示す。以下、図１５を用いて利用者の追加処理について説明する。以下の処理は、利用者が追加されるノードの経路管理部５０４が行う。なお、以下の説明において、ストレージ評価テーブル５０８に経路評価値が項目として含まれていると仮定する。

【０１４４】

まず、経路管理部５０４は、利用者が追加されるストレージセットを特定する

ストレージセット番号を取得する。このストレージセット番号は、例えば追加される利用者がアクセスした際に入力することとしても良い。

【０１４５】

経路管理部５０４は、ストレージセット管理テーブル５０９から、そのストレージセット番号に対応するストレージセット構成情報を取得する（Ｓ４１）。続いて、経路管理部５０４は、ストレージセットの決定処理を行う（Ｓ４２）。この処理は、図１３を用いて説明した処理と同様である。

【０１４６】

経路管理部５０４は、Ｓ４１で取得したストレージセット構成情報に基づいて、Ｓ４２で決定されたストレージセットを構成するノードに、データを構成するボリウムが全て格納されているか否か判定する（Ｓ４３）。例えば、データが４つのボリウムに分割されている場合、Ｓ４２でストレージセットとして４つのノードが決定されるが、これらの４つのノードに、既に４つのボリウムが格納されているか否か判定する。

【０１４７】

Ｓ４２で決定されたストレージセットを構成するノードに、データを構成するボリウムが全て格納されている場合（Ｓ４３：Ｙｅｓ）、処理を終了する。この場合、新たに追加される利用者が利用するノードにおいても、良好な利用効率が得られるからである。

【０１４８】

Ｓ４２で決定されたストレージセットを構成するノードに、データを構成するボリウムが全て格納されていない場合（Ｓ４３：Ｎｏ）、経路管理部５０４は、ストレージセット構造情報に基づいて、Ｓ４２で決定されたストレージセットを構成するノードのうちでボリウムを格納していないノード（以下、未使用ノード）のノード名及び、不足しているボリウムを格納しているノード（以下、既存ノード）のノード名を取得する（Ｓ４４）。

【０１４９】

経路管理部５０４は、ストレージ評価テーブル５０８から、未使用ノードおよび既存ノードについてのストレージ評価情報を取得し、各ストレージ評価情報に

含まれるホップ数を比較する（Ｓ４５）。未使用ノードのホップ数の方が、既存のノードのホップ数よりも小さい場合（Ｓ４５：Ｙｅｓ）、Ｓ４８に進み、そうでない場合（Ｓ４５：Ｎｏ）、Ｓ４６に進む。

【０１５０】

Ｓ４６において、経路管理部５０４は、更に、各ストレージ評価情報に含まれる経路評価値を比較する（Ｓ４６）。未使用ノードの経路評価値に定数ａ（１以上）を掛算した値の方が、既存のノードの経路評価値よりも小さい場合（Ｓ４６：Ｙｅｓ）、Ｓ４８に進み、そうでない場合（Ｓ４６：Ｎｏ）、Ｓ４７に進む。

【０１５１】

Ｓ４７において、経路管理部５０４は、更に、各ストレージ評価情報に含まれるストレージ評価値を比較する（Ｓ４６）。未使用ノードのストレージ評価値に定数ｂ（１以上）を掛算した値の方が、既存のノードのストレージ評価値よりも小さい場合（Ｓ４７：Ｙｅｓ）、Ｓ４８に進み、そうでない場合（Ｓ４７：Ｎｏ）、処理を終了する。現在の状態でも、追加される利用者にとって良好な利用効率が得られるからである。

【０１５２】

Ｓ４８において、経路管理部５０４は、不足しているボリウムを既存ノードのストレージから複製し、その複製を未使用ノードのストレージに書き込むことと決定する。この結果に基づいて、制御パケット生成部５０１は、決定結果に基づいて、既存ノード宛てに、ストレージセット番号、ボリウム番号、及び未使用ノードのノード名を含み、制御内容が「ボリウムの複写」である制御パケットを生成し、そのパケットは制御装置Ｃから送信される。このパケットに基づいて、ボリウムの複製が、未使用ノードに生成される。

【０１５３】

続いて、経路管理部５０４は、Ｓ４１で取得されたストレージセット構造情報中の未使用ノードのプロパティに、複写されたボリウムのボリウム番号、ボリウムが複写であることを示すフラグ及び使用状況情報を追記し、処理を終了する。これにより、追加される利用者にとっても良好な利用効率が得られるようになる。

【０１５４】

なお、ストレージ評価テーブル５０８に経路評価値が含まれていない場合、上記処理において、Ｓ４６は行われない。また、評価の順を、構成によって変更することとしても良い。

【０１５５】

以下、４つのボリウムにデータを複数のストレージに分散格納する場合を例として、利用者の追加処理について具体的に説明する。説明において、既存の利用者ＡはノードＡにアクセスし、追加される利用者ＥはノードＥにアクセスすると仮定する。

【０１５６】

図１６は、利用者の追加処理を説明する図である。図１６には、向かって左側に、ノードＡから見た、各ノードについてのストレージ評価値、ホップ数、各ノードのストレージに格納されているボリウムを示す表が、右側に、ノードＥから見たそれらを示す表が記載されている。図１６において、左向きの矢印は、「左の表と同じ」ことを示す。

【０１５７】

図１６に即して説明すると、図１６の左側のノードＡから見た表において、最もストレージ評価値が高い４つのノードＡ、Ｂ、Ｅ、Ｇにそれぞれボリウムａ、ｂ、ｃ及びｄが格納されている。このことから、ストレージセットはノードＡから見て利用効率が良いように最適化されていることがわかる。一方、追加される利用者ＥのノードＥから見て最もストレージ評価値が高い４つのノードは、ノードＡ、Ｂ、Ｄ及びＥである。上記の通り、ノードＡ、Ｂ及びＥには、既にボリウムａ、ｂ及びｃが格納されているが、ノードＥには、ボリウムが格納されていない。

【０１５８】

この場合、未使用ノードはノードＤとなり、既存ノードはノードＧとなり、不足するボリウムはｄとなる。

ここで、図１６によると、ノードＥからノードＤまでのホップ数は「１」であり、ノードＥからノードＧまでのホップ数は「２」である。従って、ノードＥの

経路管理部５０４は、ノードＥにボリウムｄの複製ｄ’を複写する事により、ノードＥから見ても利用効率が良いようにストレージセットを最適化することに決定する。

【０１５９】

利用者の追加処理についてさらに説明する。上記の図１６の説明において、既存の利用者ＡがノードＡにアクセスにするように最適化された広域分散ストレージシステムに、ノードＥを利用する利用者Ｅを追加する処理について説明した。次に、さらに第３の利用者としてノードＣを利用する利用者Ｃを追加する処理について説明する。図１７は、第３の利用者の追加処理を説明する図である。図１７には、向かって左側に、ノードＡから見た、各ノードについてのストレージ評価値、ホップ数、各ノードのストレージに格納されているボリウムを示す表が、中心に、ノードＥから見たそれらを示す表、右側に、ノードＣから見たそれらを示す表が記載されている。図１７において、左向きの矢印は、「左の表と同じ」ことを示す。また、この説明においても、格納されるべきデータは、４つのボリウムに分割されると仮定する。

【０１６０】

第３の利用者を追加する場合も、図１７で説明した処理と基本的に同様の処理を行う。つまり、経路管理部５０４は、ストレージセット管理テーブル５０９から、利用者が追加されることになるストレージセット番号に対応するストレージセット構成情報を取得し、ストレージセット構成情報に含まれるストレージ評価値が最も高い４つのノードを選択する。図１７によると、選択されるノードは、ノードＢ、Ｃ、Ｄ及びＥである。これらのノードがノードＣから見て、ストレージセットを構成するノードとして決定される。

【０１６１】

続いて、経路管理部５０４は、決定されたストレージセットを構成するノードに、データを構成するボリウムが全て格納されているか否か判定する。図１７によれば、ノードＢにボリウムｂ、ノードＤにボリウムｄ’（ｄの複製）、ノードＥにボリウムｃが書き込まれているが、決定されたストレージセットを構成するいずれのノードにもボリウムａは書き込まれていない事が分かる。また、未使用

ノードはノードＣであり、既存ノード、つまりボリウムａ又はａ’（ａの複製）を格納しているノードは、ノードＡ及びノードＦであることが分かる。

【０１６２】

図１７の場合、ノードＣから見た、未使用ノードと既存ノードのホップ数を比較した結果、既存ノードＡ及びＦのホップ数（それぞれ２及び３）は、未使用ノードＣのホップ数（０）よりも大きいため、経路管理部５０４は、既存ノードからボリウムａを複写して未使用ノードＣに格納することとして決定する。

【０１６３】

ここで、既存ノードはノードＡ及びノードＦの２つあるため、複写の方法は以下の２通り考えられる。いずれの方法を採用することとしても良い。

方法１）ホップ数が少なく、且つ評価値が高いノードから不足ボリウムを複写して未使用ノードに格納することとする。図１７の場合、ノードＡからボリウムａを複写する。

【０１６４】

方法２）既存のノードのうちから２以上のノードを選択し、それらのノードから不足ボリウムを分散させて読み出して複写し、未使用ノードに格納する。図１７の場合、ノードＡ及びノードＦからボリウムｃを分散させて読み出す。

【０１６５】

次に、分散配置ストレージシステムを構成する各ストレージの状態の確認方法について説明する。各ストレージの状態が正常であるか異常であるかを常時確認することが可能なように、分散配置ストレージシステムを構成することも可能である。この場合、各ノードの制御装置Ｃはローカルノードのストレージに対して状態を問い合わせ、ストレージはそれに対して正常な状態を示すキープアライブ（keep alive）信号を発信する。ストレージからのキープアライブ信号が途絶えた場合、そのノードの制御装置Ｃは、ストレージセット管理テーブル５０９を参照して、そのノードを用いて構成されているストレージセットを特定する。そして、特定されたストレージセットを構成する他のノードに対して、異常が生じたことを通知する。異常を検出した制御装置Ｃ及び異常を通知された制御装置Ｃ、つまり全ての制御装置Ｃは、ストレージセット構造情報内の全体プロパティ中の

状態情報を、異常状態を示す値「Ｒ（Ｒｅｄ）」とする。さらに、全ての制御装置Ｃは、そのストレージセットへの書き込みを遮断するように設定する。但し、利用者からのデータ読み出し要求に対しては、各制御装置Ｃは継続して処理を行う。

【０１６６】

次に、書き込みデータ及び復元データの一時保管について説明する。利用者から格納を指示されたデータは、利用者がアクセスしているローカルノードで冗長化された後に複数のボリウムに分割され、そのローカルノードからストレージセットを構成する各ノードに転送される。この際に、ローカルノードの制御装置Ｃは、そのデータのストレージセット番号と対応付けて、そのデータを一時記憶領域に保管することとしてもよい。

【０１６７】

この場合、ローカルノードの制御装置Ｃは、利用者から、ストレージセット番号とともにデータの読み出し要求を受信した際に、まず、そのストレージセット番号に対応するデータが一時記憶領域に格納されているか否か判定する。

【０１６８】

データが一時記憶領域に格納されていた場合、制御装置Ｃは、そのデータを利用者に送信する。そうでない場合、制御装置Ｃは、ストレージセット番号に対応するストレージセット構造情報をストレージセット管理テーブル５０９から取得し、そのストレージセット構造情報に基づいて各ノードからデータを復元するために必要なボリウムを取得し、それらのボリウムを用いて復元されたデータを利用者に送信する。このとき、復元されたデータは、制御装置Ｃの一時記憶領域に格納される。一時記憶領域が一杯になった場合、制御装置Ｃは、使用頻度の低いデータを削除し、そのデータに使用されていた領域を再利用する。これにより、データ読み出し時のレスポンスを向上させることが可能となる。

【０１６９】

次に、データをストレージに書き込むタイミングについて説明する。データの格納要求があった場合に、制御装置Ｃは、すぐにデータを複数のボリウムに分割して複数のノードのストレージに分散格納するのではなく、一旦、データを一時

記憶領域に格納した後に、ストレージに分散格納することとしても良い。

【０１７０】

より具体的には、制御装置Ｃは、利用者からデータの格納要求を所定回数受信するのを待ち、その間に格納要求されたデータを一時記憶領域に格納する。データの格納要求を所定回数受信した場合、制御装置Ｃは、一時記憶領域に格納された各データを複数のボリウムに分割して、ボリウムの１つをローカルノードのストレージに格納させ、他のボリウムそれぞれについては、ストレージセットを構成する他のノードに転送し、各ノードのストレージに格納させる。その後、一時格納領域に書きこまれたデータを消去する。これにより、ローカルノード以外の他のノードにボリウムを転送する回数を低減する事が可能となるため、トラフィックの効率を上げることが可能となる。

【０１７１】

なお、上記において、制御装置Ｃは、データの格納要求を所定回数受信するまでデータを一時記憶領域に格納するとして説明したが、データの格納要求を所定回数受信するまで待つ代わりに、所定時間、データを一時記憶領域に格納することとしてもよい。この場合も、上記と同様の効果を得ることが可能である。

【０１７２】

次に、広域分散ストレージシステムを構成するストレージ上に格納されたボリウムを更新する場合の処理について説明する。上記のように、ストレージ上には、データが複数のボリウムに分割されて格納されている。このようなデータを更新する際に、マルチキャストパケットを利用する事としても良い。以下、この場合の処理について説明する。

【０１７３】

１）まず、ストレージセットを構成するノードを、ボリウムごとにグループ化する。例えば、図１７の右側の表に示すようなストレージセットの場合、ストレージセット管理部５０５は、ボリウムａのグループとしてノードＡ、Ｃ及びＦ、ボリウムｂのグループとしてノードＢ、ボリウムｃのグループとしてノードＥ、ボリウムｄのグループとしてノードＤ及びＧを定義する。このグループは、不図示のグループテーブルに格納することとしても良い。

【０１７４】

２）続いて、利用者が各ボリウムを更新する際には、利用者が利用するローカルノードからマルチキャストパケットを用いて、ボリウムａ、ｂ、ｃ及びｄの各グループに分けられたノードのストレージに対して書き込みを行う。

【０１７５】

３）ローカルノードのストレージセット管理部５０５は、各ボリウムのグループに分けられたノードから正常に更新が完了した旨の通知を受けると、更新が完了したこととする。

【０１７６】

４）更新の際に異常が発生した場合、ローカルノードのストレージセット管理部５０５は、異常が発生したノード内のボリウムに対して再度の更新処理を行う。

【０１７７】

さらに、ストレージ上に分割して格納されたデータが利用者によって更新される場合、その更新処理の間、他の利用者がそのデータを構成するボリウムの原本或いは複製を更新することを禁止することとしても良い。これにより、更新されるデータについて原本と複製の内容の統一を図ることが可能となる。以下、この場合の処理について説明する。

【０１７８】

行われる処理の手順の概要は以下のとおりである。

１）まず、利用者はノードにアクセスし、更新したいデータのストレージセット番号を指定し、更新要求をそのノードに送信する。以下、この利用者がアクセスするノードをローカルノードという。ローカルノードの制御装置Ｃ内のストレージセット管理部５０５は、その利用者に対してロックキーを発行する。ロックキーは、データの更新を要求した利用者を識別するために使用され、広域分散ストレージシステムの利用者及びセッションごとにユニークなものである。図２２に、このロックキーの機能を追加したアクセス管理テーブル５１０及びローカルボリウム管理テーブル５１１の一例を示す。

【０１７９】

２）ローカルノードの制御装置Ｃ内のストレージセット管理部５０５は、そのストレージセット番号によって特定されるボリウムを他の利用者が更新する事を禁止し（以下、ロックという）、さらに、そのストレージセットを構成する他のノードに、対応するボリウムをロックするよう依頼する。

【０１８０】

３）依頼を受信した各ノードの制御装置Ｃ内のストレージセット管理部５０５は、各ボリウムをロックする。

４）ローカルノードの制御装置Ｃ内のストレージセット管理部５０５は、各ボリウムがロックされていることを確認する。

【０１８１】

５）ローカルノードの制御装置Ｃは、更新するべきデータを冗長化して複数のボリウムに分割し、各ボリウムを、それぞれを格納するべきノードに送信する。

６）ボリウムの送信及び更新が終了すると、各ノードの制御装置Ｃ内のストレージセット管理部５０５は、ロックを解除する。

【０１８２】

以下、上記の手順２）から６）を順に詳しく説明する。まず、手順２）について、図１８を用いて詳しく説明する。図１８に示す手順は、利用者からストレージセット番号の指定と、更新要求を受信したローカルノードの制御装置Ｃ内のストレージセット管理部５０５によって行われる。

【０１８３】

まず、ストレージセット管理部５０５は、利用者から受信したストレージセット番号に対応するアクセス管理情報を取得し、そのアクセス管理情報に基づいて、更新要求が出されたアクセス単位（論理ブロック）の状態が、「ロック状態」であるのか否か判定する（Ｓ５１）。そのアクセス単位の状態が「ロック状態」であると判定された場合（Ｓ５１：Ｙｅｓ）。Ｓ５２に進み、そのアクセス単位の状態が「ロック状態」以外の状態である場合、Ｓ５７に進む。Ｓ５７において、ストレージセット管理部５０５は、ロックに失敗した旨を利用者に通知し、処理を終了する（終了パターン２：異常終了）。異常終了の場合、更新を行う事はできない。

【０１８４】

Ｓ５２において、ストレージセット管理部５０５は、ロックキーを生成する。

続いて、ストレージセット管理部５０５は、ストレージセット番号に対応するストレージセットについてのストレージセット構造情報をストレージセット管理テーブル５０９から取得し、そのストレージセットを構成するノードに、そのストレージセットをロックする依頼を通知する（Ｓ５３）。なお、ロック依頼には、ストレージセット番号、ストレージセットアクセス番号及びロックキーが含まれる。なお、ストレージセットを構成するノードにローカルノードが含まれる場合もあることをいうまでもない。

【０１８５】

通知を受けた各ノードは、ロック処理を行う（Ｓ５４）。この処理については後述する。

さらに、ローカルノードのストレージセット管理部５０５は、Ｓ５３で通知した各ノード全てからロックが完了した旨の通知を待つ（Ｓ５５）。全ノードからロックが完了した旨の通知を受信した場合（Ｓ５５：Ｙｅｓ）、Ｓ５８に進む。そうでない場合（Ｓ５５：Ｎｏ）、Ｓ５６に進む。

【０１８６】

Ｓ５６において、ストレージセット管理部５０５は、利用者に対してロックに失敗した事を通知し、処理を終了する（終了パターン２：異常終了）。

Ｓ５８において、ストレージセット管理部５０５は、更新要求が出されたアクセス単位についてのアクセス管理情報にふくまれるプロパティを「ロック状態」に更新し、さらに、Ｓ５２で生成されたロックキーを追加するように、アクセス管理テーブル５１０を更新する（Ｓ５８）。さらに、ストレージセット管理部５０５は、その利用者に対してロックキーを発行して（Ｓ５９）、処理を終了する（終了パターン１：正常終了）。

【０１８７】

続いて、手順３）について、図１９を用いて説明する。この手順３）は、図１８のＳ５４において、ロックの依頼を受信した各ノードの制御装置Ｃによって行われる処理に相当する。

【０１８８】

まず、ローカルボリウム管理部５０６は、ロック依頼と共に受信したストレージセット番号及びストレージセットアクセス番号に対応するボリウム管理情報をローカルボリウム管理テーブル５１１から取得し、そのボリウム管理情報に基づいて、更新要求が出されたアクセス単位（論理ブロック）の状態が、「ロック状態」であるのか否か判定する（Ｓ６１）。そのアクセス単位の状態が「ロック状態」であると判定された場合（Ｓ６１：Ｙｅｓ）。Ｓ６２に進み、そのアクセス単位の状態が「ロック状態」以外の状態である場合、Ｓ６６に進む。Ｓ６６において、ローカルボリウム管理部５０６は、ロックに失敗した旨を利用者に通知し、処理を終了する（終了パターン２：異常終了）。異常終了の場合、更新を行う事はできない。

【０１８９】

Ｓ６２において、ローカルボリウム管理部５０６は、更新要求が出されたアクセス単位についてのボリウム管理情報にふくまれるプロパティに「ロック状態」を示すフラグを追加し、さらにロックキーを追加する。続いて、ストレージセット管理部５０５は、ロックが完了した旨を、ロック依頼を通知してきたノードの制御装置Ｃに通知する（Ｓ６３）。

【０１９０】

さらに、ストレージセット管理部５０５は、アクセス管理テーブル５１０に、ロック対象となっている論理ブロックについてのアクセス管理情報が格納されているか否か判定する（Ｓ６４）。アクセス管理情報が格納されている場合（Ｓ６５：Ｙｅｓ）、ストレージセット管理部５０５は、そのアクセス管理情報のプロパティを「ロック状態」に更新する。アクセス管理情報がアクセス管理テーブル５１０に格納されていない場合（Ｓ６５：Ｎｏ）、処理を終了する（終了パターン１：正常終了）。

【０１９１】

続いて、手順４）から６）までについて図２０及び２１を用いて説明する。まず、更新要求を出した利用者は、発行されたロックキーとともに、更新したいデータの内容をローカルノードに送信する（不図示）。まず、図２０について説明

する。図２０に示す処理は、ローカルノードの制御装置Ｃによって行われる。

【０１９２】

ローカルノードのストレージセット管理部５０５は、更新対象となる論理ブロック（ロック対象の論理ブロックでもある）についてのアクセス管理情報をアクセス管理テーブル５１０から取得する。（Ｓ７１）。続いて、ストレージ管理部５０５は、Ｓ７１で取得したアクセス管理情報に含まれるプロパティに基づいて、更新対象となるブロックはロックされているか否か判定する（Ｓ７２）。更新対象となるブロックがロックされている場合（Ｓ７２：Ｙｅｓ）、Ｓ７３に進み、更新対象となるブロックがロックされていない場合（Ｓ７２：Ｎｏ）、Ｓ７８に進む。

【０１９３】

Ｓ７３において、ストレージセット管理部５０５は、利用者から受信したロックキーが、アクセス管理情報に含まれるロックキーと一致するか否か判定する。２つのロックキーが位置しない場合（Ｓ７３：Ｎｏ）、ストレージセット管理部５０５は、利用者に対し、書き込み（更新）に失敗した旨を通知し（Ｓ７９）、処理を終了する（終了パターン２：異常終了）。

【０１９４】

２つのロックキーが一致した場合（Ｓ７３：Ｙｅｓ）、データ変換部３は、利用者から取得したデータを冗長化し、複数のボリウムに分割する。さらに、ストレージセット制御部５０１、ストレージセット管理テーブル５０９から、アクセス対象となっているストレージセットについてのストレージセット構造情報を取得する。経路管理部５０４は、ストレージセット管理テーブル５０９からストレージセット番号に対応するストレージセット構造情報を取得し、そのストレージセット構造情報に基づいてそのストレージセットを構成する各ノードに、書き込み依頼と共にボリウムを送信するように制御情報及び経路情報を生成し、パケット生成部４は、制御情報及び経路情報を付加したパケットを各ノードに送信する。なお、書き込み依頼にはストレージセット番号、ストレージセットアクセス番号及びロックキーが含まれる。これを受けて、各ノードでは、データをストレージＳに書き込む処理が行われる（Ｓ７４）。この処理については図２１を用いて

後述する。

【０１９５】

続いて、ストレージセット管理部５０５は、書き込み依頼を送信した全てのノードから書き込み完了の通知を受信するまで待つ（Ｓ７５）。全てのノードから書き込み完了の通知を受信しなかった場合（Ｓ７５：Ｎｏ）、ローカルノードの制御装置Ｃは、そのストレージセットを構成する各ノードに、書き込み依頼と共にボリウムを再び送信する。これを受けて、各ノードでは、データをストレージＳに書き込む処理が再び行われる（Ｓ８０）。この処理は、Ｓ７４と同じである。

【０１９６】

全てのノードから書き込み完了の通知を受信した場合（Ｓ７５：Ｙｅｓ）、制御装置Ｃは、書き込みの要求を出した利用者に書き込み完了の通知を送信する（Ｓ７６）。ストレージセット管理部５０５は、更新対象となる論理ブロックについてのアクセス管理情報をアクセス管理テーブル５１０から取得し、アクセス管理情報に含まれるプロパティから「ロック状態」を示すフラグを削除し（Ｓ７７）、処理を終了する（終了パターン１：正常終了）。

【０１９７】

Ｓ７８において、制御装置Ｃはロック処理を行う。ロック処理について、上記において図１８及び１９を用いて既に説明したため、ここでは説明をしない。Ｓ７８のロック処理が正常終了した場合（終了パターン１）、Ｓ７４に進み、Ｓ７８のロック処理が異常終了した場合（終了パターン２）、Ｓ７９に進む。

【０１９８】

次に、図２１について説明する。図２１の処理は、上記の図２０のＳ７４の処理に相当し、書き込み依頼を受信した各ノードの制御装置Ｃによって行われる。

まず、ローカルボリウム管理部５０６は、ローカルボリウム管理テーブル５１１から書き込み依頼と共に受信したストレージセット番号及びストレージセットアクセス番号に対応するボリウム管理情報をローカルボリウム管理テーブル５１１から取得する（Ｓ８１）。続いて、ローカルボリウム管理部５０６は、そのボリウム管理情報に含まれるロックキーが、書き込み依頼に含まれるロックキーと

一致するか否か判定する（Ｓ８２）。２つのロックキーが一致する場合（Ｓ８２：Ｙｅｓ）、Ｓ８３に進み、一致しない場合（Ｓ８２：Ｎｏ）、ローカルボリウム管理部５０６は、書き込みに失敗した旨を、書き込み依頼を送信してきたノードの制御装置Ｃに送信し、処理を終了する（終了パターン２：異常終了）。

【０１９９】

Ｓ８３において、パケット解析部７は、書き込み依頼と共に受信したパケットからボリウムを取出して、ストレージＩＦ８を介してそのボリウムをストレージＳに書き込む。続いて、制御パケット生成部５０２は、書き込みが完了した旨の通知を、書き込み依頼を送信してきたノードの制御装置Ｃに送信する（Ｓ８４）。続いて、ローカルボリウム管理部５０６は、Ｓ８１で取得したボリウム管理情報に含まれるプロパティから、「ロック状態」を示すフラグ及びロックキーを削除し（Ｓ８５）、処理を終了する。

【０２００】

次に、上記のロック処理において、更新の際にマルチキャストパケットを用いる場合について説明する。この場合の処理の手順の概要は以下のとおりである。なお、上記のマルチキャストを用いる更新処理の場合と同様に、処理の前にストレージセットを構成するノードを、ボリウムごとにグループ化することが必要である。各ノードの制御装置Ｃには、各ボリウムのグループを構成するノードとそのグループの代表ノードを示す情報を含む不図示のノードグループテーブルが備えられている。

【０２０１】

１）ストレージセット中で、各ボリウムを格納するノードを代表する代表ノード（例えば、原本を格納するノード）に対して、データ更新の前にデータアクセス単位でロック処理を行う。

【０２０２】

２）原本は同じグループに属するノードに対して、上記更新要求を送信した利用者を確認できるようにして（ロックキーを使用）ロック処理を行う。なお、１）及び２）の手順は、上記のロック処理と同様である。

【０２０３】

３）更新要求を送信した利用者は、マルチキャストパケットを用いて各ボリウムの更新内容を送信する。

４）更新内容を含むパケットを受信した各ノードは、ロックキーを用いて更新要求を送信した利用者と、そのパケットを送信した利用者とが同一であることを確認し、確認できた場合、ボリウムの更新を行う。

【０２０４】

５）上記４）における更新の際、ボリウムのグループの代表ノードは、更新完了を示すパケットを利用者がアクセスしているノードに送信する。代表ノード以外のノードは、そのグループの代表ノードに更新完了を示すパケットを送信する。

【０２０５】

６）代表ノードは、自ノードに格納されたボリウムに対するロックを解除する。また、代表ノードは、自グループに属する他のノードから更新完了を示すパケットを受信した場合、そのノードに格納されたボリウムに対するロックを解除する。

【０２０６】

７）代表ノードは、自グループに属する他のノードのうちで、更新完了を示すパケットを送信してこないノードがある場合、そのノードに対して更新処理を実行する。

【０２０７】

以下、図２３から図２５を用いて、上記３）からの手順についてより詳しく説明する。なお、図２３から図２５に示す処理は、図２０及び図２１に示す処理と同様の手順を含むため、図２３から図２５において図２０及び図２１と同様の手順には同じ番号を付し、説明を省略する。まず、図２３を用いて、更新要求を出した利用者がアクセスするノードで行われる処理について説明する。以下において、利用者がアクセスするノードをローカルノードという。

【０２０８】

図２３では、図２０のＳ７４からＳ７７及びＳ８０の代わりに、Ｓ９１からＳ９５を行う点が図２０に示す処理と異なる。以下、Ｓ９１からＳ９５について説

明する。なお、Ｓ９１からＳ９５は、上記3）から7）に示す手順のうち、ローカルノードの制御装置Ｃによって行われる手順を示す。

【０２０９】

Ｓ７１からＳ７３の後、経路管理部５０４は、ストレージセット管理テーブル５０９からストレージセット番号に対応するストレージセット構造情報を取得し、そのストレージセット構造情報に基づいてそのストレージセットを構成する各ノードに、書き込み依頼と共にそのノードに書き込むべきボリウムを送信するように制御情報及び経路情報を生成し、パケット生成部４は、制御情報及び経路情報を付加したパケットを各ノードにマルチキャストで送信する。なお、書き込み依頼にはストレージセット番号、ストレージセットアクセス番号及びロックキーが含まれる。これを受けて、各ノードでは、ボリウムをストレージＳに書き込む処理が行われる（Ｓ９１）。この処理については図２４及び図２５を用いて後述する。

【０２１０】

続いて、ストレージセット管理部５０５は、書き込み依頼を送信したノードのうち、代表ノードから書き込み完了の通知を受信するまで待つ（Ｓ９２）。ここで、ストレージセット管理部５０５は、不図示のノードグループテーブルに基づいて、全ての代表ノードから書き込み完了の通知を受信したか否か判定する。なお、図２３から図２５において原本を格納するノードを代表ノードと仮定している。

【０２１１】

全ての代表ノードから書き込み完了の通知を受信しなかった場合（Ｓ９２：Ｎｏ）、ローカルノードの制御装置Ｃは、その代表ノードに、書き込み依頼と共に書き込むべきボリウムを再び送信する。これを受けて、各ノードでは、データをストレージＳに書き込み処理が再び行われる（Ｓ８０）。この処理は、Ｓ９１と同じである。

【０２１２】

全てのノードから書き込み完了の通知を受信した場合（Ｓ９２：Ｙｅｓ）、ローカルノードのストレージセット管理部５０５は、書き込みが完了した旨を利用

者に通知する（Ｓ９３）。続いて、ストレージセット管理部５０５は、更新対象となる論理ブロックについてのアクセス管理情報をアクセス管理テーブル５１０から取得し、アクセス管理情報に含まれるプロパティから「ロック状態」を示すフラグを削除し、処理を終了する（終了パターン１：正常終了）。

【０２１３】

また、ローカルノードの制御装置Ｃは、Ｓ７８及びＳ７９も行う。Ｓ７８及びＳ７９の処理については図２０と同様であるため説明を省略する。

次に、図２４を用いて、図２３のＳ９１において行われる処理について説明する。図２４の処理は、各ボリウムのグループにおける代表ノードによって行われる。

【０２１４】

図２４では、図２１のＳ８５の後に、更に、Ｓ１０１及びＳ１０２を行う点が図２１に示す処理と異なる。以下、Ｓ１０１及びＳ１０２について説明する。

Ｓ８１からＳ８５の後、代表ノードのストレージセット管理部５０５は、その代表ノードが属するグループ内の他のノードの全てから書き込み完了の通知を受信するまで待つ（Ｓ１０１）。この代表ノードはノードグループテーブル（不図示）に基づいて、Ｓ１０１の判定を行う。全てのノードから書き込み完了の通知を受信しなかった場合（Ｓ１０１：Ｎｏ）、代表ノードの制御装置Ｃは、書き込み完了の通知を送信してこなかったノード対するボリウムの書き込みを、そのノードに代わって実行し（Ｓ１０２）、Ｓ１０１に戻る。

【０２１５】

全てのノードから書き込み完了の通知を受信した場合（Ｓ１０１：Ｙｅｓ）、処理を正常終了する（終了パターン１）。

次に、図２５を用いて、図２３のＳ９１において行われる処理について説明する。図２５の処理は、ストレージセットを構成するノードのうちで、代表ノード以外のノードによって行われる。

【０２１６】

図２５では、図２１のＳ８３の後に、Ｓ８４の代わりに、Ｓ１１１を行う点が図２１に示す処理と異なる。以下、Ｓ１１１について説明する。

Ｓ８１からＳ８３の後、代表ノード以外のノードのストレージセット管理部５０５は、そのノードが属するグループの代表ノードへ書き込み完了の通知送信する（Ｓ１１１）。

【０２１７】

次に、新規のボリウムの作成手順について説明する。新規のボリウムは、たとえば、ボリウムの複製を作成する際や障害からの復旧の際等において作成される。

【０２１８】

以下、ボリウムの複製を作成する際を例にとって、新規のボリウムの作成手順について説明する。図２６は、向かって左から順に、ノードＡにアクセスする利用者Ａ、ノードＢにアクセスする利用者Ｂ及びノードＣにアクセスする利用者が、広域分散ストレージシステムを利用する場合に、それぞれのノードから見た、ストレージ評価値、ホップ数及び各ノードのストレージに格納されているボリウムを示す表である。以下、図２６を用いて、複製の作成方法の決定について説明する。なお、この説明において、データは４つのボリウムに分割されていると仮定する。

【０２１９】

図２６によると、利用者Ｃが使用するべきストレージセットは、ノードＢ、Ｃ、Ｄ及びＥであり、利用者Ｃを広域分散ストレージシステムに追加する際に、ノードＣにボリウムａの複製が作成される。このボリウムａの複製は、以下の２通りの方法で作成することができる。

【０２２０】

１）ノードＡ又はＦにのいずれかに格納されたボリウムａから作成する。

２）ボリウムａの複製を他のボリウムｂ、ｃ及びｄから冗長を用いて再現する。

【０２２１】

複製が作成されるボリウムを格納するノードと、その他のボリウムを格納するノードのノードＣから見たストレージ評価値を比較し、複製が作成されるボリウムを格納するノードの最高ストレージ評価値を、他の各ボリウムを格納するノー

ドの最高ストレージ評価値のそれぞれが上回っていた場合、上記のうち２）の方法を採用し、そうでない場合、１）の方法を採用する。

【０２２２】

例えば、図２６の場合、上位４つのストレージ評価値は以下のようになる。

ボリウムａを格納するノードの最高ストレージ評価値：１１．３（ノードＡ）

ボリウムｂを格納するノードの最高ストレージ評価値：１７．０（ノードＢ）

ボリウムｃを格納するノードの最高ストレージ評価値：１４．８（ノードＥ）

ボリウムｄを格納するノードの最高ストレージ評価値：２１．８（ノードＤ）

ボリウムａを格納するノードＡのストレージ評価値は、他のボリウムを格納するノードのストレージ評価値のいずれよりも低い。従って、この場合、ノードＣの経路管理部５０４は、ノードＣ作成されるボリウムａの複製を、それぞれノードＢ、Ｄ及びＥに格納されるボリウムｂ、ｄ及びｃから冗長を利用して再現することと決定し、この決定に基づいて、各ノードからボリウムｂ、ｃ及びｄはノードＣに転送され、ノードＣのパケット解析部７は、受信したボリウムｂ、ｃ、ｄからボリウムａを再現して、ストレージＩＦ８を介してストレージＳにボリウムａを書き込む。

【０２２３】

次に、広域分散ストレージシステムを構成するノードの一部に障害が発生した場合の処理について説明する。まず、ノードに障害が発生した場合に、残っているノードから見て最適なストレージセットを再設定する処理について説明する。説明において、データは４つのボリウムに分割されていると仮定する。また、障害発生前後の各ノードの状態を図２７（ａ）及び（ｂ）に示すような状態として仮定する。なお、図２７（ａ）及び（ｂ）において、向かって左側に、ノードＡから見た、各ノードについてのストレージ評価値、ホップ数、各ノードのストレージに格納されているボリウムを示す表が、中心に、ノードＥから見たそれらを示す表、右側に、ノードＣから見たそれらを示す表が記載されている。図２７（ａ）及び（ｂ）において、左向きの矢印は、「左の表と同じ」ことを示す。

【０２２４】

まず、図２７（ａ）に示すような状態において、ノードＡの利用者Ａは、ノー

ドＡ、Ｂ、Ｅ及びＧからそれぞれボリウムａ、ｂ、ｃ及びｄを取得する。ノードＥの利用者Ｅは、ノードＡ、Ｂ、Ｄ及びＥからそれぞれボリウムａ、ｂ、ｄ’及びｃを取得する。ノードＣの利用者Ｃは、ノードＢ、Ｃ、Ｄ及びＥからそれぞれボリウムｂ、ａ’、ｄ’及びｃを取得する。なお、「’」は、そのボリウムが複製である事を示す。

【０２２５】

このような状態で、ノードＡのストレージに障害が発生した場合、利用者Ａ及び利用者Ｅは、ノードＡからボリウムａを取得する事ができなくなる。この場合、ボリウムａは、ノードＣ及Ｆにも存在するため、ノードＡ及びノードＥのストレージセット管理部は、ノードＡの代わりにノードＣ又はＦのいずれかからボリウムａを取得する事に決定する。

【０２２６】

どのノードからボリウムａを取得するのか決定する方法は、基本的に、ストレージセットの決定処理と同様に、以下のように決定する。

１）ストレージ評価値が高い方を採用する。

【０２２７】

２）ストレージ評価値が同じ場合にホップ数が少ない方を採用する。

図２７（ｂ）において、ノードＡから見て、ノードＣとノードＦのストレージ評価値は、それぞれ、「１０．８」及び「８．３」である。従って、ノードＡのストレージセット管理部５０５は、ノードＡのストレージの代わりにノードＣのストレージからボリウムａを取得する事として決定する。同様に、ノードＥから見て、ノードＣとノードＦのストレージ評価値はそれぞれ、「１６．３」及び「１３．０」である。従って、ノードＥのストレージセット管理部５０５は、ノードＡのストレージの代わりにノードＣのストレージからボリウムａを取得する事として決定する。

【０２２８】

さらに、障害が生じたノードＡのストレージに格納されていたボリウムａは原本であったが、今回の障害の結果、ボリウムａの原本が存在しない事となってしまうため、ノードＣとノードＦのいずれかを代表ノードとして扱う事と決定する

。なお、図２７（ｂ）においては、各ノードからのノードＣ及びノードＦに対するストレージ評価値のうち、より平均値が大きいノードＣを代表ノードとしている。

【０２２９】

さらに、障害からの復旧の際の、新規ボリウムの作成手順について説明する。図２８は、図２７と同様に、向かって左から順に、ノードＡにアクセスする利用者Ａ、ノードＥにアクセスする利用者Ｅ及びノードＣにアクセスする利用者Ｃが、広域分散ストレージシステムを利用する場合に、それぞれのノードから見た、ストレージ評価値、ホップ数及び各ノードのストレージに格納されているボリウムを示す表である。以下、図２８を用いてノードＡのストレージが障害から普及した場合、ノードＡのストレージにボリウムを格納する際に、どのボリウムの複製をどうやって作成するのか、その決定手順について説明する。なお、この説明においても、データは４つのボリウムに分割されていると仮定する。

【０２３０】

１）まず、ストレージセット管理部５０５は、復旧するノードのストレージ評価値は、復旧前に使用されていたノードのストレージ評価値よりも高いか否か判定する。復旧するノードのストレージ評価値の方が復旧前に使用されていたノードのストレージ評価値よりも高い場合、ストレージセット管理部５０５は、復旧するノードのストレージＳにボリウムの複製を書き込むことと決定する。複製されるボリウムは、復旧前に使用されていたノードのうち、最低のストレージ評価値を持つノードのストレージに格納されていたボリウムとする。

【０２３１】

例えば、図２８に示すように、ノードＡでは、ノードＡのストレージＳの復旧前に、ノードＢ、Ｃ、Ｅ及びＧのストレージが使用されていたが、ノードＡのストレージ評価値の方がこれらより高い。また、ノードＢ、Ｃ、Ｅ及びＧのうち最低のストレージ評価値を持つノードはノードＣであり、そのノードＣのストレージＳに格納されているボリウムはボリウムａである。同様に、ノードＥでは、ノードＡのストレージＳの復旧前に、ノードＢ、Ｃ、Ｄ及びＥのストレージが使用されていたが、このうちの最低のストレージ評価値を持つノードＣよりもノード

Ａのストレージ評価値の方が高い。ノードＣでは、ノードＡの復旧前後において、追加されるボリウムを格納するノードのストレージ評価値は、使用するノードの変更を要するほど高くないため、何もしない。

【０２３２】

２）複製の作成は、最もストレージ評価値の高いノードで行われる。通常、ボリウムの複製が書き込まれるストレージに直接接続されているノードで行われる。図２８の場合、ノードＡ（つまり、複製が作成されるストレージに接続されているノード）の制御装置Ｃは、ノードＡのストレージＳにボリウムａの複製を書き込む。

【０２３３】

３）複製を作成する際、ノードＡのストレージセット管理部５０５は、複製される元となるボリウムを格納するノードのストレージ評価値と、その他のボリウムを格納するノードのストレージ評価値とを比較し、比較結果に基づいて、ストレージに格納されたボリウムを元にして複製を作成するか、それとも他のボリウムから冗長を利用してボリウムを再現するか、決定する。この決定方法は、上記の複製の作成方法と同様であるため、詳しい説明は省略する。

【０２３４】

図２８の場合、ボリウムａを格納するノードのうち最大のストレージ評価値を持つノードはノードＣであり、その値は１０．８であり、この値は、他のボリウムを格納するノードのストレージ評価値よりも低い、従って、ストレージセット管理部５０５は、他のボリウムｂ、ｃ及びｄから冗長を利用してボリウムａを再現する事と決定する。

【０２３５】

次に、ストレージセットの管理について説明する。ノードの追加や削除を繰り替えすと、利用されないボリウム等が存在するようになる。このような場合、利用されないボリウムを削除することにより、ストレージの利用効率を向上させるようにストレージセットを管理ことが可能である。以下、ストレージセットの管理について説明する。

【０２３６】

以下において、データは４つのボリウムに分割され、ノードの追加や削除の結果、図２９に示すようなボリウム構成になっていると仮定する。なお、図２９（ａ）及び（ｂ）は、それぞれ向かって左から順に、ノードＡにアクセスする利用者Ａ、ノードＥにアクセスする利用者Ｅ及びノードＣにアクセスする利用者Ｃが、広域分散ストレージシステムを利用する場合に、それぞれのノードから見た、ストレージ評価値、ホップ数及び各ノードのストレージに格納されているボリウムを示す表であり、これは、図２８と同様である。ここで、図２９（ａ）は、使用されていないボリウムの削除前の状態を示し、図２９（ｂ）は、使用されていないボリウムの削除後の状態を示す。

【０２３７】

１）一定時間経過ごと、或いは、各ノードのストレージＳに格納されるボリウムの状態が変更されるごとに、各ノードにおいて使用されるストレージセットを示す情報を各ノードで交換する。あるいは、任意の１つのノードに、その情報を集める。

【０２３８】

２）交換された情報に基づいて、どのノードによっても使用されていないボリウムが合った場合、１つのノードのストレージセット管理部５０５は、そのノードのボリウムを削除する事と決定し、そのノードに対してそのボリウムを削除するよう指示する制御パケットを送信する。

【０２３９】

例えば、図２９において、ノードＡにおいてストレージセットとして使用されているノードはノードＡ、Ｂ、Ｅ及びＧであり、ノードＥにおいてストレージセットとして使用されているノードはノードＡ、Ｂ、Ｄ及びＥであり、ノードＣにおいてストレージセットとして使用されているノードは、ノードＢ、Ｃ、Ｄ及びＥである。（使用されているノードは最もストレージ評価値が高い４ノードである）　従って、ノードＦは、どのノードの利用者によっても使用されていないことがわかる。従って、ノードＦのストレージＳに格納されるボリウムａ’は削除される（図２９（ｂ）参照）。

【０２４０】

次に、データの複製の再生又はデータの再生を逐次に行う処理について説明する。データの複製又は再生は、利用者が追加された場合や新たなノードが追加された場合等において行われるが、緊急に行う必要が無いことも多い。この場合、データの複製又は再生は、ネットワークの空き時間等を利用して逐次に行われることとしてもよい。これにより、トラフィックの有効利用を可能とする。以下、この場合の処理について説明する。以下の処理は、利用者からデータの読み込み要求又は書き込み要求を受信したノード（以下、ローカルノード）によって行われる。

【０２４１】

図３０に、逐次にデータの複製の作成又は再生を行う場合に、処理についてのフローチャートを示す。図３０に示すように、まず、利用者は、読み出したいデータのストレージセット番号を指定して読み込み要求又は書き込み要求をアクセス先となっているノードに送信する。ローカルノードのストレージセット管理部５０５は、ストレージセット管理テーブル５０９からストレージセット番号に対応するストレージセット構造情報を取得し、そのストレージセット構造情報に基づいてそのローカルノードが使用するストレージセットに、データを構成する全てのボリウムが格納されているか否か判定する（Ｓ１２１）。

【０２４２】

ローカルノードが使用するストレージセットに、データを構成する全てのボリウムが格納されている場合（Ｓ１２１：Ｙｅｓ）、データの読み込み・書き込み処理が行われる（Ｓ１２６）。この読み込み・書き込み処理については、既に説明した。

【０２４３】

ローカルノードが使用するストレージセットに、データを構成する全てのボリウムが格納されていない場合（Ｓ１２１：Ｎｏ）、ストレージセット管理部５０５は、読み出し要求を受信した場合は、各ノードから取得したボリウムから冗長化によって要求されたデータを生成する。書き込み要求を受信した場合は、ストレージセット管理部５０５は、受信したデータを冗長化し、複数のボリウムに分割して各ノードに書き込む。その際に、ストレージセット管理部５０５は、読み

出し又は書き込みアクセス管理テーブル５１０から該当するストレージセット番号を持つアクセス管理情報を取得し、そのアクセス管理情報に含まれるそのボリウムへのアクセスに関するプロパティに、アクセスされたデータが、ストレージＳから読み出された完全データであるのか、冗長化を利用して生成された生成データであるのかを示すフラグを付す（Ｓ１２２）。

【０２４４】

続いて、利用者から読み出し要求を受信した場合、ストレージセット管理部５０５は、Ｓ１２１で取得したストレージセット構造情報に基づいて、ストレージセットを構成するノードのうちでボリウムを格納していないノード（不完全ノード）を特定し、そのノードに格納するためのボリウムを、ストレージセットを構成する他のノードから読み出されたボリウムから生成する（Ｓ１２３）。

【０２４５】

ローカルストレージの制御装置Ｃは、生成されたボリウムを書き込むように指示して、不完全ノードのストレージＳにそのボリウムを転送する。これを受けて、不完全ノードでは、ボリウムの書き込み処理が行われる（Ｓ１２４）。この書き込み処理は、逐次にネットワークの空き時間を利用して行われる。なお、ストレージセット管理部５０５は、このボリウムが完全データでないことを示すフラグをＳ１２１で取得したストレージセット構造情報に含まれる不完全ノードについてのプロパティに付す。

【０２４６】

逐次書き込みにおいて、その書き込みによるアクセスを示すアクセス管理情報中のプロパティに、アクセスされたデータが、ストレージＳから読み出された完全データであるのか、冗長化を利用して生成された生成データであるのかを示すフラグを付される。逐次書き込みが完了すると、アクセス管理情報中のプロパティには、生成データであることを示すフラグは付されないことになる。なお、２度目以降の逐次書き込みにおいて、ローカルノードの制御装置Ｃは、ストレージセット構造情報及びアクセス管理情報内の生成データを示すフラグに基づいて、不完全ノードと、それに格納すべきボリウムを特定することができる。

【０２４７】

書き込み処理が終了すると、ストレージセット管理部５０５は、アクセス管理テーブル５１０を参照し、ストレージセット番号に対応するアクセス管理情報に、生成データであることを示すフラグが付されているか否か判定する。生成データであることを示すフラグが付されたアクセス管理情報が無ければ、ローカルノードが使用するストレージセットに、データを構成する全てのボリウムが格納されていることとなる（Ｓ１２５：Ｙｅｓ）。この場合、ローカルノードのストレージセット管理部５０５は、このボリウムが完全データでないことを示すフラグをストレージセット構造情報に含まれる不完全ノードについてのプロパティから取り除く（Ｓ１２７）。生成データであることを示すフラグが付されたアクセス管理情報がある場合、まだ逐次書き込みが終了していない（Ｓ１２５：Ｎｏ）。この場合、一旦処理を終了し、Ｓ１２５を繰り返す。Ｓ１２５の判定がＹｅｓとなるまで、又は、一定回数Ｓ１２５を繰り返すまで、Ｓ１２５は繰り返されることとしてもよい。

【０２４８】

図３１は、それぞれ向かって左から順に、ノードＡにアクセスする利用者Ａ、ノードＥにアクセスする利用者Ｅ及びノードＣにアクセスする利用者Ｃが、広域分散ストレージシステムを利用する場合に、それぞれのノードから見た、ストレージ評価値、ホップ数及び各ノードのストレージに格納されているボリウムを示す表である。以下、図３１を用いてデータの複製又は再生を逐次に行う場合について具体的に説明する。この図３１は、ノードＣにアクセスする利用者が追加された際の状態を示す。なお、以下の説明においてデータは４つのボリウムに分割されて分散して格納されると仮定する。また、以下の説明において、利用者Ｃがアクセスするノードをローカルノードという。

【０２４９】

まず、図３１に示すように、利用者Ｃが使用するストレージセット（つまり、ノードＣから見たストレージセット）は、ノードＢ、Ｃ、Ｄ及びＥである。ノードＤ以外のノードには既にボリウムが格納されている。不足しているボリウムはボリウムｄである。しかし、ノードＢ、Ｃ及びＥに格納されているボリウムｂ、ａ及びｃから、データを復元することは可能であるため、ノードＤへのボリウム

ｄの書き込みを即時に行う事は不要である。

【０２５０】

そこで、ノードＤのストレージセット管理部５０５は、アクセス管理テーブル５１０から該当するストレージセット番号を持つアクセス管理情報を取得し、そのアクセス管理情報において、そのボリウムｄへのアクセスに関するプロパティに、アクセスされるデータが、ストレージＳから読み出された完全データであるのか、冗長化を利用して生成された生成データであるのかを示すフラグを付す。

【０２５１】

利用者Ｃからデータの読み出し要求を受信した場合、ローカルノードの制御装置Ｃは、ノードＢ、Ｃ及びＥからそれぞれボリウムｂ、ａ及びｃを受信し、それらのボリウムから冗長化を利用してデータを再現して利用者に転送する。その際に、ローカルノードの制御装置Ｃはボリウムｄを作成し、ノードＤのストレージＳにそのボリウムｄを逐次に格納させる。

【０２５２】

図３２に、制御装置の配置方法の変形例を示す。上記説明において、制御装置Ｃは各ノードに備えられるとした。しかし、制御装置Ｃは、利用者の端末に備えられる事としても良い。この場合、利用者の端末が、データを複数のボリウムに分割し、各ボリウムの格納先となるストレージを選択してデータを書き込む。さらに、利用者の端末が、複数のボリウムを各ストレージから読み出してデータを復元する。この場合でも、上記の広域分散ストレージシステムと同様の効果を得ることができる。図３２には、ノードＡを利用する利用者Ａの端末が、データを３つのボリウムに分割し、ノードＡ、Ｂ及びＧの３つのノードのストレージＳに格納させる場合を示す。この場合、利用者Ａの端末は、データを読み出す際には、ノードＡ、Ｂ及びＧから３つのボリウムを読み出してデータを復元する。なお、この場合でも、上記の変形例に対応して様々な変形例が考え得る。

【０２５３】

上記において説明した制御装置Ｃ内の制御部５は、コンピュータを用いて構成することができる。コンピュータは、少なくとも、ＣＰＵとそのＣＰＵに接続されたメモリを備え、さらに、外部記憶装置、媒体駆動装置を備えてもよい。それ

らはバスにより互いに接続されている。

【０２５４】

メモリは、例えば、ＲＯＭ（Read Only Memory）、ＲＡＭ（Random Access Memory）等であり、処理に用いられるプログラムとデータを格納する。制御部５を構成する各部及び各テーブルは、コンピュータのメモリの特定のプログラムコードセグメントにプログラムとして格納される。なお、制御装置Ｃによって行われる処理は、図を用いて既に説明した。

【０２５５】

ＣＰＵは、メモリを利用して上述のプログラムを実行することにより、必要な処理を行う。

外部記憶装置は、例えば、磁気ディスク装置、光ディスク装置、光磁気ディスク装置等である。外部記憶装置は、各テーブルを実現する。さらに、上述のプログラムをコンピュータの外部記憶装置に保存しておき、必要に応じて、それらをメモリにロードして使用することもできる。

【０２５６】

媒体駆動装置は、可搬記録媒体を駆動し、その記録内容にアクセスする。可搬記録媒体としては、メモリカード、メモリスティック、フレキシブルディスク、ＣＤ－ＲＯＭ（Compact Disc Read Only Memory）、光ディスク、光磁気ディスク、ＤＶＤ（Digital Versatile Disk）等、任意のコンピュータ読み取り可能な記録媒体が用いられる。この可搬記録媒体に上述のプログラムを格納しておき、必要に応じて、それをコンピュータのメモリにロードして使用することもできる。

【０２５７】

また、ネットワークＩＦを介して、上述のプログラムを、ネットワークＩＦを介して、プログラムをダウンロードすることとしてもよい。

以上、本発明の実施形態について説明したが、本発明は上述した実施形態及び変形例に限定されるものではなく、その他の様々な変更が可能である。

【０２５８】

（付記１）　コンピュータが、データを冗長化して複数のボリウムに分割し、

各ボリウムを、ネットワークを介して分散配置された複数のストレージに分散して格納するデータ格納方法であって、

帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離に基づいて、前記分散配置された各ストレージについて利用対象としての望ましさを示す評価値を算出し、

前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択する、

ことを特徴とするデータ格納方法。

【０２５９】

（付記２）　前記評価値の算出において、前記書き込みを依頼するノードから各ストレージまでのホップ数をさらに用いる、

ことを特徴とする付記１記載のデータ格納方法。

【０２６０】

（付記３）　前記システムの利用者に対して、前記ストレージセットを仮想的な１つのストレージとして提供する、

ことを更に含むことを特徴とする付記１に記載のデータ格納方法。

【０２６１】

（付記４）　前記データを前記ストレージセットから読み込む際には、前記ストレージセットに書きこまれた前記複数のボリウムのうち冗長化部分を含まないボリウムを各ストレージから読み出し、

前記読み出されたボリウムを用いて前記データを復元する、

ことを更に含むことを特徴とする付記１に記載のデータ格納方法。

【０２６２】

（付記５）　前記データを読み込む際には、各ストレージについて、前記帯域幅及び前記コストに基づいてレスポンスの良さを示す利用優先度を算出し、

前記利用優先度に基づいて、冗長化部分を含まないボリウムとして、前記複数のボリウムのうちいずれのボリウムを各ストレージから読み出すか決定する、

ことを更に含むことを特徴とする付記３に記載のデータ格納方法。

【０２６３】

（付記６）　前記ストレージセットとして選択されなかったストレージに、前記複数のボリウムうちの第１のボリウムの複製を格納する、

ことを更に含むことを特徴とする付記１に記載のデータ格納方法。

【０２６４】

（付記７）　前記第１のボリウムの複製を作成する際に、前記評価値に基づいて、前記第１のボリウムを格納するストレージから前記第１のボリウムを複写するのか、前記複数のボリウムのうちの前記第１のボリウム以外のボリウムから冗長を利用して前記第１のボリウムを再現するのか、２つの作成方法のうちのいずれかを選択する、

ことを更に含むことを特徴とする付記６に記載のデータ格納方法。

【０２６５】

（付記８）　前記評価値に基づいて、前記ストレージセットとして選択されなかったストレージの中から前記ボリウムの複製を格納するストレージを選択する、

ことを更に含むことを特徴とする付記６に記載のデータ格納方法。

【０２６６】

（付記９）　同一のボリウムを格納するべき複数のストレージに対して、マルチキャストでボリウムを書き込む、

ことを更に含むことを特徴とする付記６に記載のデータ格納方法。

【０２６７】

（付記１０）　前記第１のボリウムの複製をストレージに書き込む際に、多数回に分けて書き込み処理を行う、

ことを特徴とする付記６に記載のデータ格納方法。

【０２６８】

（付記１１）　前記ストレージセットのうちの第１のストレージに障害が発生した場合、前記ストレージセットのうちの他のストレージへの書き込みを制限する、

ことを更に含むことを特徴とする付記１に記載のデータ格納方法。

【０２６９】

（付記１２）　前記ストレージセットのうち第３のストレージに障害が発生した場合、前記評価値に基づいて、前記ストレージセットとして選択されているストレージ以外の第４のストレージを、前記第３のストレージの代わりに選択する

、

ことを更に含むことを特徴とする付記１に記載のデータ格納方法。

【０２７０】

（付記１３）　前記ストレージセットの選択後、一定のタイミングで、各ノードにおけるストレージセットを再選択し、

再選択の結果、どのノードからも利用されていないボリウムがあった場合、該ボリウムをストレージから削除する、

ことを更に含むことを特徴とする付記１に記載のデータ格納方法。

【０２７１】

（付記１４）　前記一定のタイミングとは、前回の選択から一定期間後、又はボリウムの状態が変更される毎である、

ことを特徴とする付記１３に記載のデータ格納方法。

【０２７２】

（付記１５）　前記データを読み込んだ後に、前記データを一定期間、任意の１つのストレージ内に一時格納し、

前記一定期間内にデータの読み出しを行う際には、一時格納されたデータを前記１つのストレージから読み出す、

ことを更に含むことを特徴とする付記１に記載のデータ格納方法。

【０２７３】

（付記１６）　一定期間内に書き込み要求されたデータを一時格納領域に一時格納し、

前記一定期間経過後に前記一時格納領域からデータを取出し、

該データを複数のボリウムに分割し、

該複数のボリウムを前記ストレージセットに書き込む、

ことを更に含むことを特徴とする付記１に記載のデータ格納方法。

【０２７４】

（付記１７）　前記一時格納したデータを含むデータに対し、読み出し又は書き込みを行う際に、前記一時格納したデータを含まない部分のデータについてのみ、読み出し又は書き込むを行う、

ことを更に含むことを特徴とする付記１５又は１６に記載のデータ格納方法。

【０２７５】

（付記１８）　前記ストレージセットに前記複数のボリウムを書き込む際に、前記書き込みを依頼するノードは、書き込みが終了するまで前記ストレージセットへの書き込み処理を禁止する、

ことを更に含むことを特徴とする付記１に記載のデータ格納方法。

【０２７６】

（付記１９）　同一のボリウムを格納するべき複数のストレージの中から、１つのストレージを代表ストレージとして決定することを更に含み、

前記複数のストレージへの書き込み処理の禁止において、

前記代表ストレージへの書き込み処理の禁止は、前記書き込みを依頼するノードによって行われ、

前記代表ストレージ以外のストレージへの書き込み処理の禁止は、前記代表ストレージによって行われる、

ことを特徴とする付記１８に記載のデータ格納方法。

【０２７７】

（付記２０）　前記代表ストレージは、原本となるボリウムを格納するべきストレージである、

ことを特徴とする付記１９に記載のデータ格納方法。

【０２７８】

（付記２１）　ネットワークを介して分散配置されたストレージを備えるシステムにデータを冗長化して複数のボリウムに分割し、各ボリウムを複数のストレージに分散して格納する制御をコンピュータに行われるコンピュータ・プログラムであって、

帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離に基づいて、前記分散配置された各ストレージについて利用対象としての望

ましさを示す評価値を算出し、

前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択する、

ことを含む制御を前記コンピュータに行わせることを特徴とするコンピュータ・プログラム。

【０２７９】

（付記２２）　ネットワークを介して分散配置されたストレージを備えるシステムにデータを冗長化して複数のボリウムに分割し、各ボリウムを複数のストレージに分散して格納する制御をコンピュータに行わせるプログラムを記録した記録媒体であって、

帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離に基づいて、前記分散配置された各ストレージについて利用対象としての望ましさを示す評価値を算出し、

前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択する、

ことを含む制御を前記コンピュータに行わせるプログラムを記録した記録媒体。

【０２８０】

（付記２３）　ネットワークを介して分散配置されたストレージを備えるシステムに使用される、データを冗長化して複数のボリウムに分割し、各ボリウムを複数のストレージに分散して格納する制御を行う制御装置であって、

帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離に基づいて、前記分散配置された各ストレージについて利用対象としての望ましさを示す評価値を算出する経路管理手段と、

前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択するストレージセット管理手段と、

を備えることを特徴とする制御装置。

【０２８１】

【発明の効果】

以上詳細に説明したように、本発明によれば、データを冗長化して複数のボリウムに分割し、各ボリウムを複数のストレージに分散して格納することによりデータのセキュリティを向上させつつも、さらに、帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離を考慮して、そのノードから見て最適なストレージを選択することにより、回線効率と災害時のデータの安全性も向上させることが可能となる。

【図面の簡単な説明】

【図１】

広域分散ストレージシステムの構成図である。

【図２】

制御装置の構成図である。

【図３】

制御装置の詳細構成図である。

【図４】

具体的な広域分散ストレージシステムの構成例を示す図である。

【図５】

経路評価テーブルの一例をす図である。

【図６】

ストレージ評価テーブルの一例を示す図である。

【図７】

ストレージセット管理テーブルの一例を示す図である。

【図８】

アクセス管理テーブルの一例を示す図である。

【図９】

ローカルボリウム管理テーブル一例を示す図である。

【図１０】

広域分散ストレージシステムにおけるデータの流れを説明する図である。

【図１１】

利用優先度及び評価値の計算処理を示すフローチャートである。

【図１２】

データ復元時にボリウムを読み出すべきストレージの決定方法を説明する図である。

【図１３】

ストレージセット管理テーブルの更新処理を示すフローチャートである。

【図１４】

冗長化したデータの分散書き込み先となるストレージの決定方法を説明する図である。

【図１５】

ノードへの利用者の追加処理を示すフローチャートである。

【図１６】

冗長化したデータの複製を格納するストレージの決定方法を説明する図である。

【図１７】

利用可能な複数のストレージから最適なストレージセットを選択する処理を説明する図である。

【図１８】

ロック処理を示すフローチャート（その１）である。

【図１９】

ロック処理を示すフローチャート（その２）である。

【図２０】

書き込み処理を示すフローチャート（その１）である。

【図２１】

書き込み処理を示すフローチャート（その２）である。

【図２２】

ストレージ内のデータ更新の際のロック処理を説明する図である。

【図２３】

マルチキャストパケットを用いた書き込み処理を示すフローチャート（その１）である。

【図２４】

マルチキャストパケットを用いた書き込み処理を示すフローチャート（その２）である。

【図２５】

マルチキャストパケットを用いた書き込み処理を示すフローチャート（その３）である。

【図２６】

ボリウムの複製の作成方法を選択する処理を説明する図である。

【図２７】

一部のストレージに障害が発生した場合に、残りのストレージの中から最適なストレージを選択しなおす処理を説明する図である。

【図２８】

ストレージが障害から復旧した際にボリウムの複製の作成方法を選択する処理を説明する図である。

【図２９】

不要となったボリウムを削除する処理を説明する図である。

【図３０】

逐次にデータを書き込み又は再生する処理を示すフローチャートである。

【図３１】

データの複製の作成又は再生を逐次に行う場合の処理を説明する図である。

【図３２】

制御装置の機能を利用者端末が備える場合を説明する図である。

【図３３】

従来の技術に係わる広域分散ストレージシステムの構成図である。

【符号の説明】

１　ユーザインタフェース（受信側）

２　ユーザインタフェース（送信側）

３　データ変換部

４　パケット生成部

５　制御部

６　データ組立部

７、３０１　パケット解析部

８　ストレージインタフェース

９　ネットワークインタフェース（送信側）

１０　ネットワークインタフェース（受信側）

３０２　データ分割部

３０３、６０３　パリティ計算部

４０１　データ管理情報付加部

４０２　制御／経路情報付加部

４０３　データ転送部

４０４　転送パケット構築部

５０１　ストレージ制御部

５０２　制御パケット生成部

５０３　ネットワーク制御部

５０４　経路管理部

５０５　ストレージセット管理部

５０６　ローカルボリウム管理部

５０７　経路評価テーブル

５０８　ストレージ評価テーブル

５０９　ストレージセット管理テーブル

５１０　アクセス管理テーブル

５１１　ローカルボリウム管理テーブル

６０１　パケット構築部

６０２　データ組立部

Ｃ　制御装置

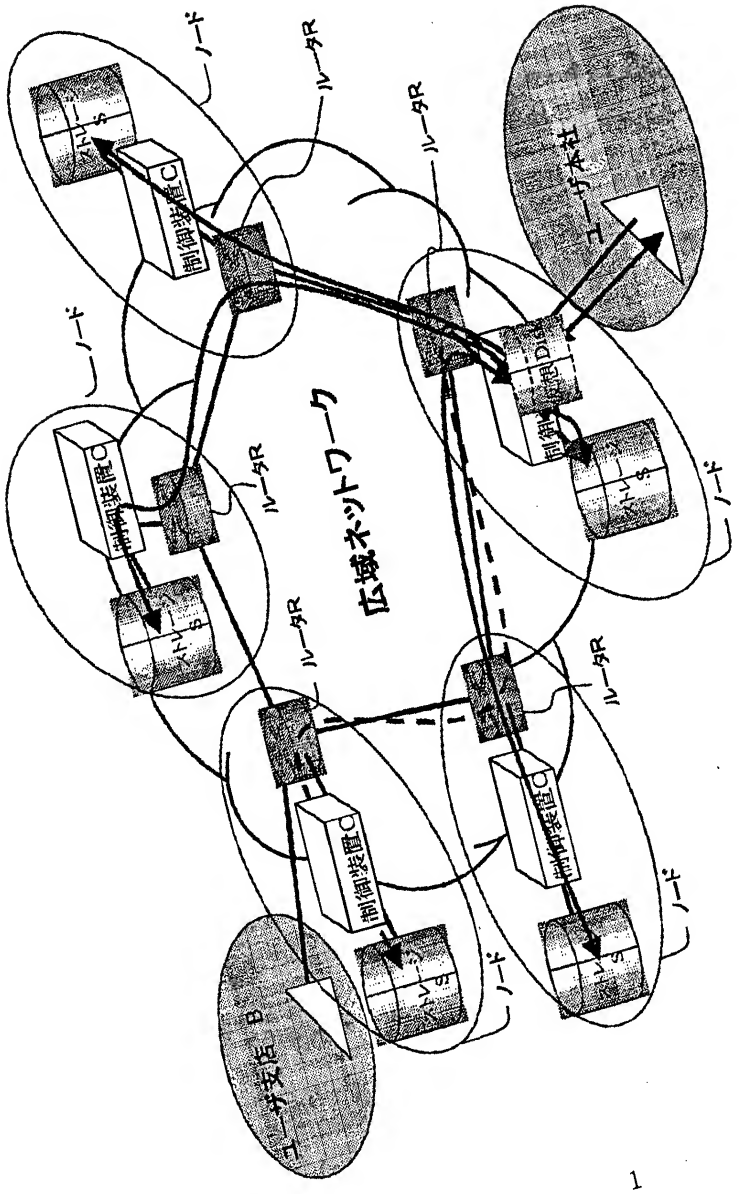
Ｒ　ルータ

Ｓ　ストレージ

【書類名】　図面

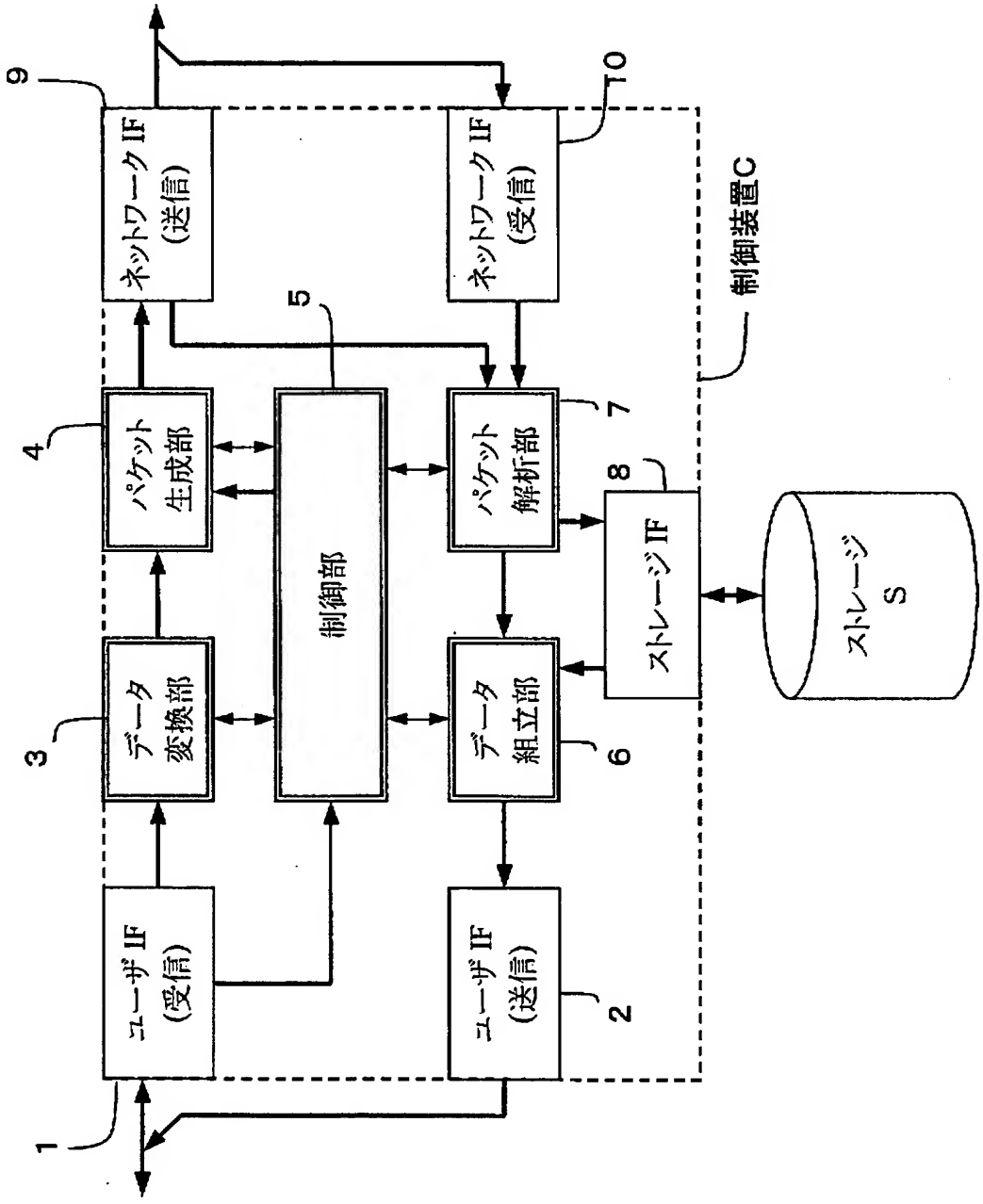
【図１】

広域分散ストレージシステムの構成図



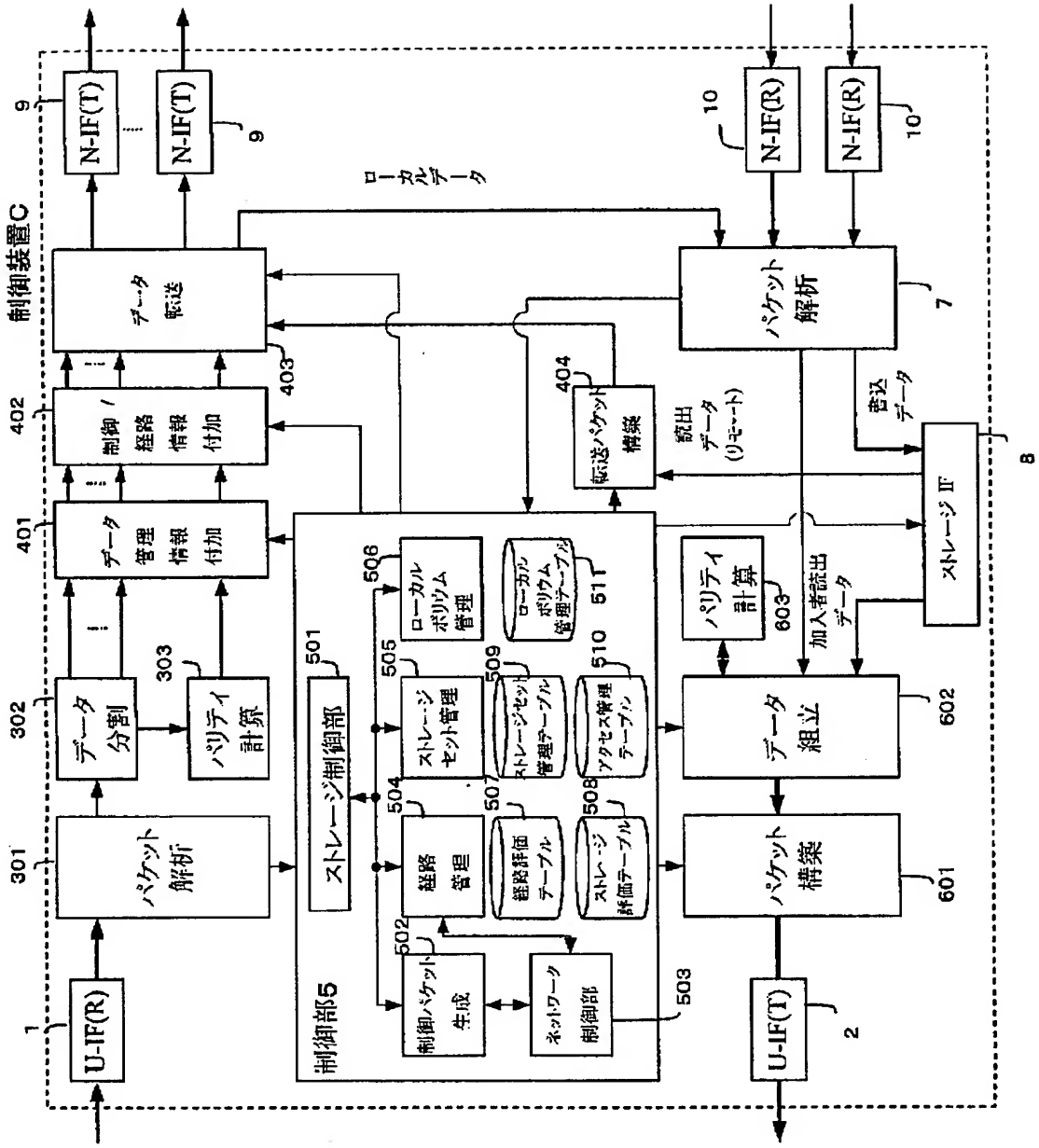
【図２】

制御装置の構成図


ネットワークIF (送信)
ネットワークIF (受信)
制御装置C
パケット 生成部
制御部
パケット 解析部
ストレージIF
ストレージ S
データ 変換部
データ 組立部
ユーザIF (受信)
ユーザIF (送信)
1
2
3
4
5
6
7
8
9
10

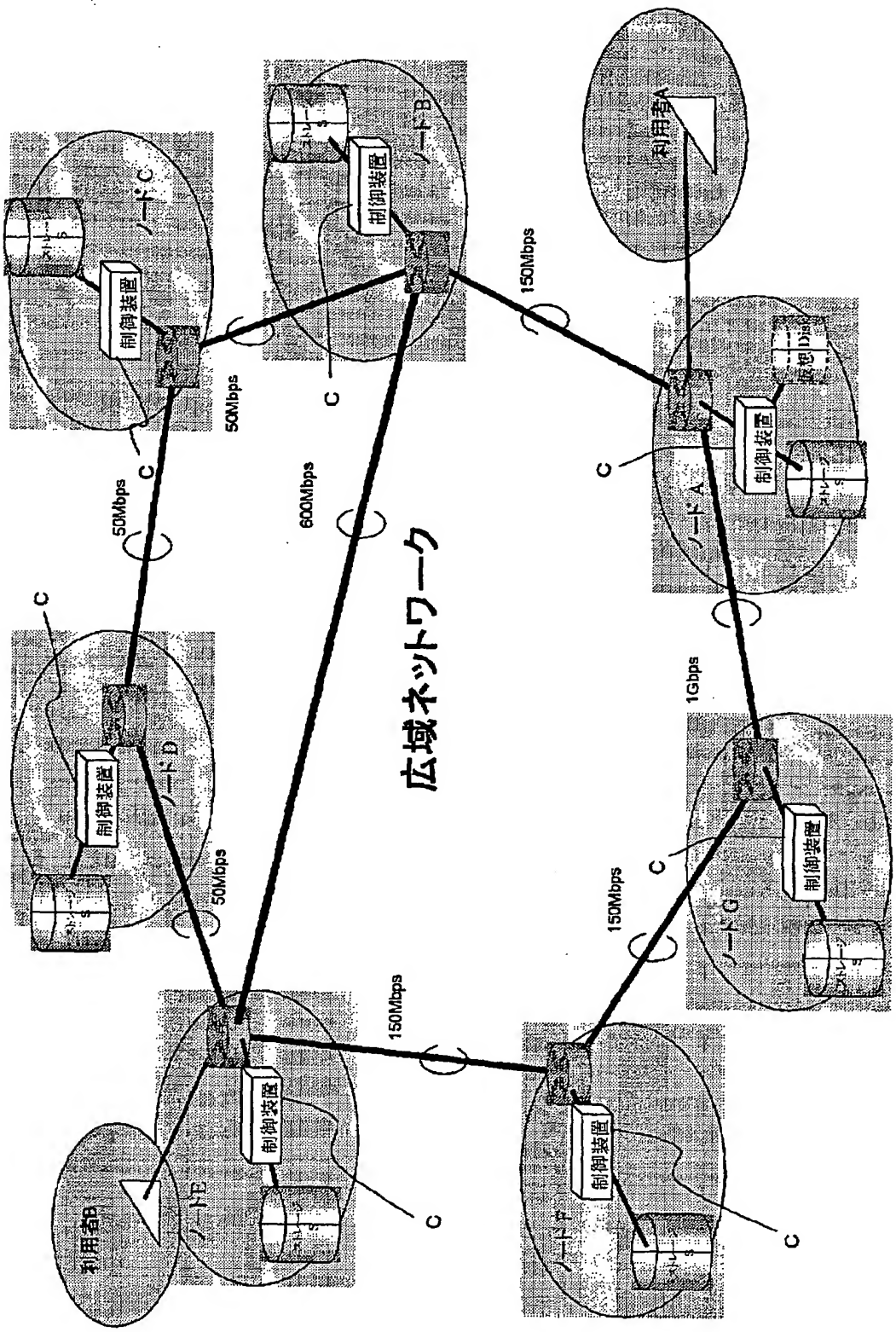
【図３】

制御装置の詳細構成図



【図４】

具体的な広域分散ストレージシステムの構成例を示す図



【図５】

経路評価テーブルの一例を示す図

経路評価テーブル

区間	帯域幅	コスト	ディスタンス	ストレージの利用優先度
local	-	-	-	∞
A-B	150	100	80	11
B-C	50	50	150	17
C-D	50	100	200	21
D-E	50	50	150	17
E-F	150	100	100	13
F-G	150	100	80	11
A-G	1000	200	10	11
B-E	600	300	200	24

【図６】

ストレージ評価テーブルの一例を示す図

ノード	経路	ストレージ評価値	ホップ数
A	local	∞	0
B	A-B	11.0	1
C	A-B-C	9.8	2
D	A-B-C-D	8.8	3
	A-B-E-D	9.6	3
E	A-B-E	11.5	2
F	A-G-F	8.3	2
G	A-G	11.0	1

【図７】

ストレージセット管理テーブルの一例を示す図

ストレージセット番号	プロパティ																							
	全体			ノードA			ノードB			ノードC			ノードD			ノードE			ノードF			ノードG		
	読出	書込	状態	分割	原本	使用	分割	原本	使用	分割	原本	使用	分割	原本	使用	分割	原本	使用	分割	原本	使用	分割	原本	使用
00000001	3	4	G	1	O	RW	3	O	RW	2	O	-W				0	O	RW				2	C	R
00000002	3	3	G	2	O	Rw				1	O	RW	0	O	RW									
00000003	4	5	R	0	O	RW	1	O	RW	4	O	RW	3	O	RW	2	O	-W						
00000004	2	3	G			RW	2	O		0	O	-W				1	O	RW						

【図８】

アクセス管理テーブルの一例を示す図

ストレージセット番号	000010001
ストレージセットアクセス番号	プロパティ
00000001	RWLO
00000002	RWLO
00000003	-W-O
00000004	-W-O
00000005	---O
00000006	-W-O

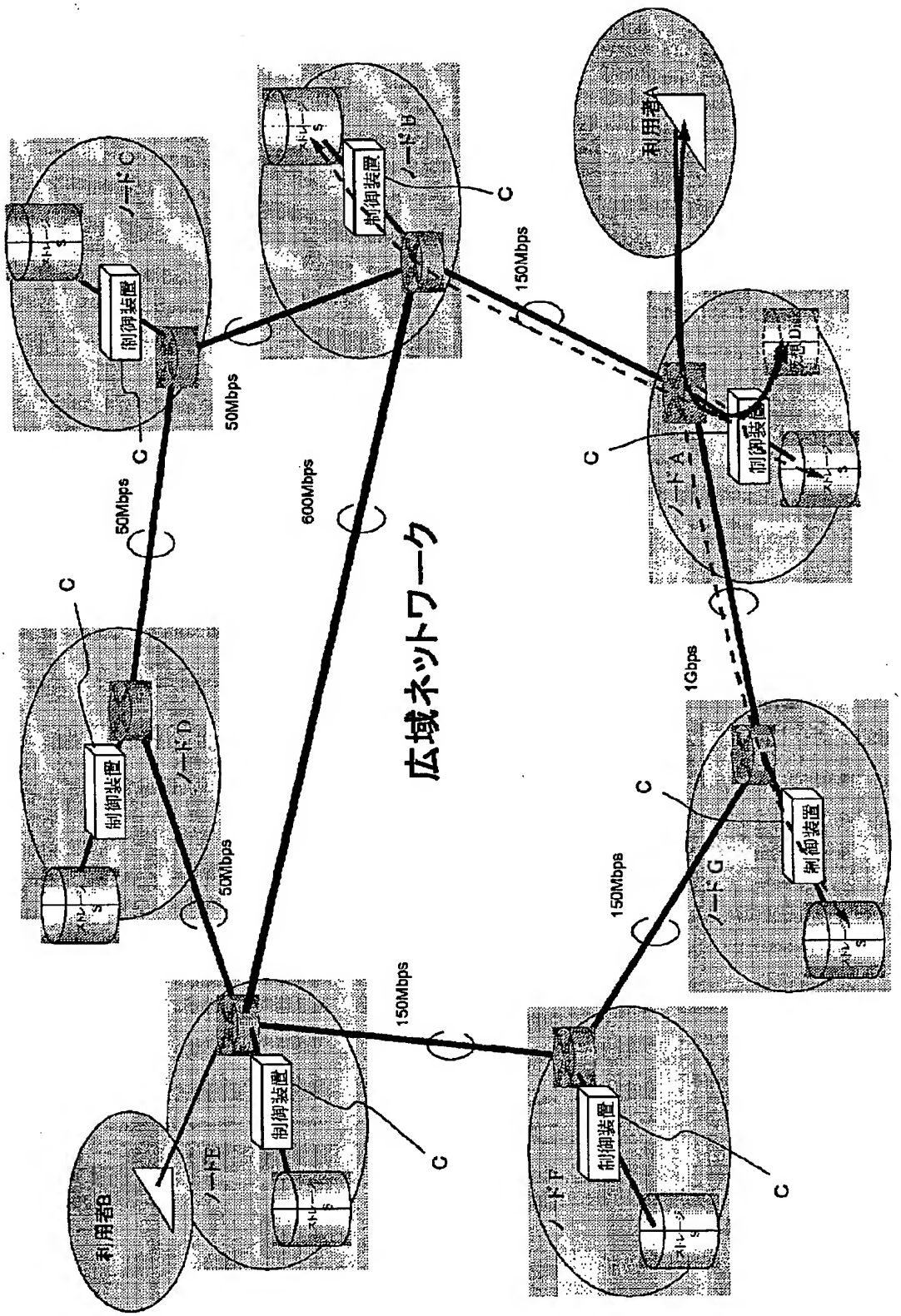
【図９】

ローカルボリウム管理テーブルの一例を示す図

ローカルストレージ		ストレージセット管理情報	
ストレージアクセス番号	プロパティ	ストレージセット管理番号	ストレージセットアクセス番号
00000001	RWL	00010021	00040001
00000002	RW-	02003001	00040004
00000003	RW-	00000000	00000000
00000004	R-F	01002111	01000100
00000005	RW-	00010021	00001562
00000006	RW-	00010021	00001563
⋮	⋮	⋮	⋮

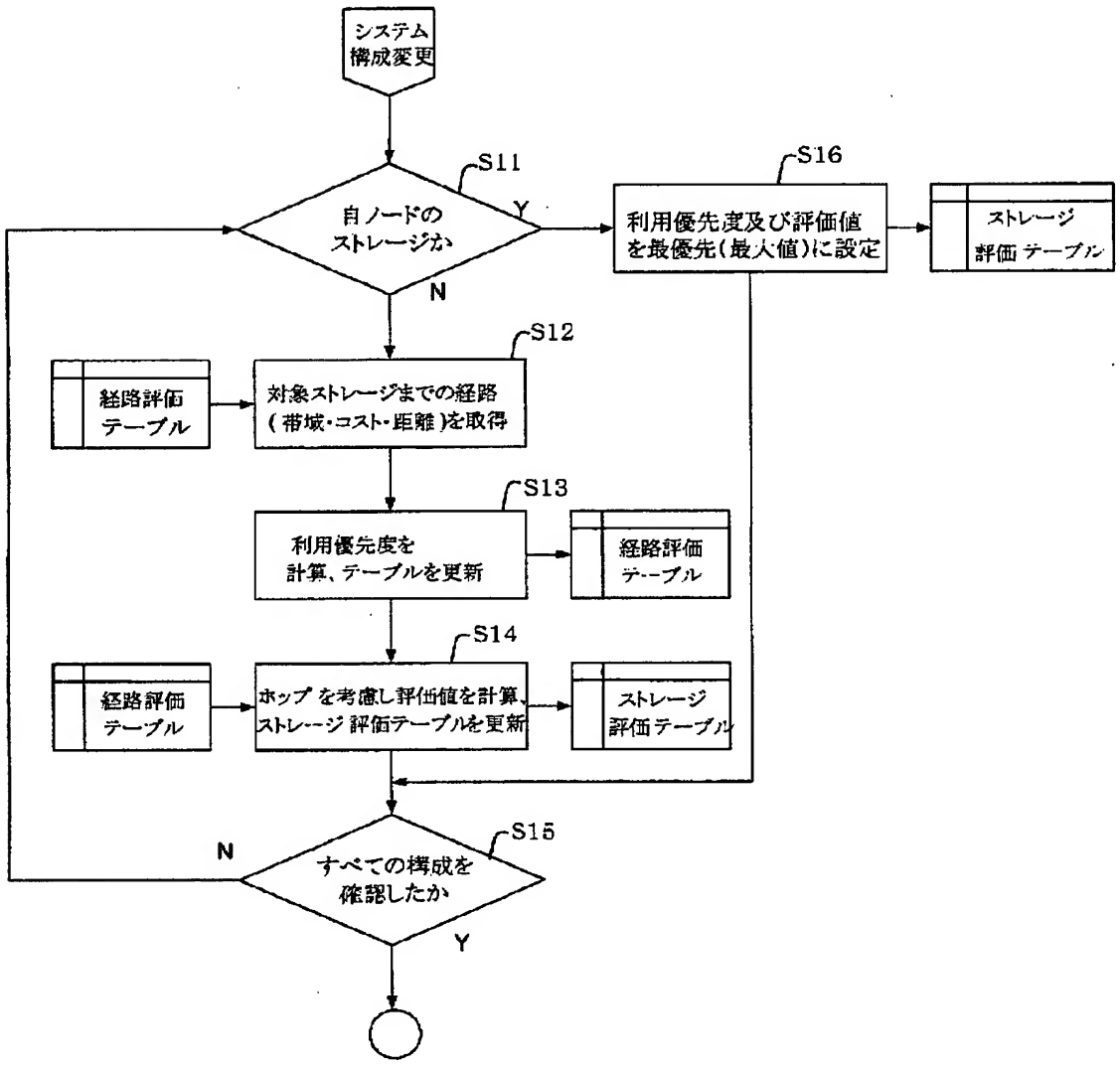
【図１０】

広域分散ストレージシステムにおけるデータの流れを説明する図



【図１１】

利用優先度の計算処理を示すフローチャート



【図１２】

データ復元時に、読み出すべき ボリウムの決定方法を説明する図

経路評価テーブル

区間	帯域幅	コスト	ディスタンス	区間の利用優先度	ストレージの利用優先度
local	-	-	-	∞	∞
A-B	150	100	80	3	11
B-C	50	50	150	2	17
C-D	50	100	200	1	21
D-E	50	50	150	2	17
E-F	150	100	100	3	13
F-G	150	100	80	3	11
A-G	1000	200	10	10	11
B-E	600	300	200	4	24

(a)

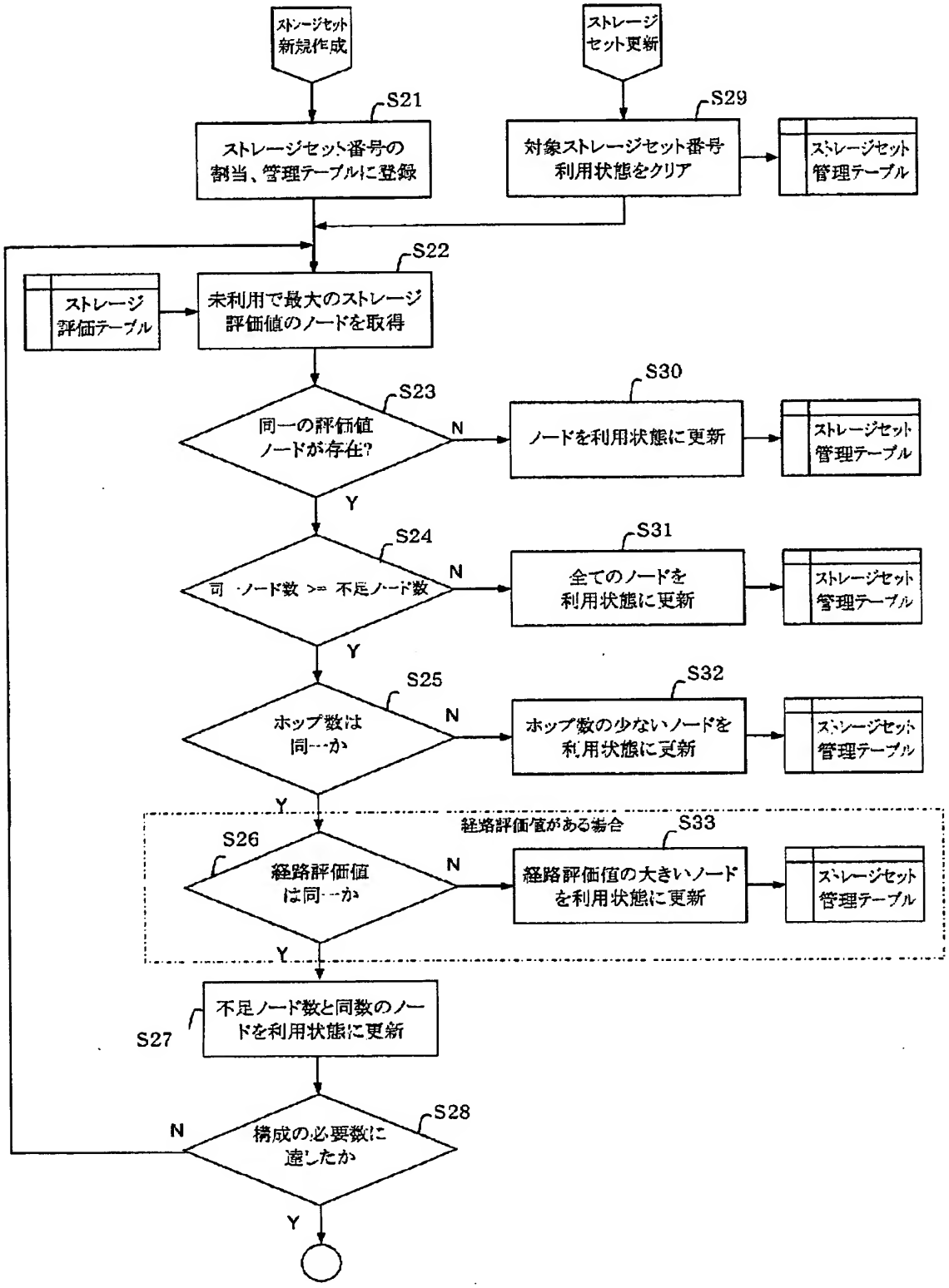
ストレージ評価テーブル

ノード	経路	経路評価値	ストレージ評価値	ホップ数
A	local	∞	∞	0
B	A-B	3	11.0	1
C	A-B-C	2.5	9.8	2
D	A-B-C-D	2	8.8	3
	A-B-E-D	3	9.6	3
E	A-B-E	3.5	11.5	2
F	A-G-F	6.5	8.3	2
G	A-G	10	11.0	1

(b)

【図１３】

ストレージセット管理テーブルの更新処理を示すフローチャート



【図１４】

冗長化したデータの分散書き込み先となる
ストレージの決定方法を説明する図

経路評価テーブル

区間	帯域[Mbps]	コスト	ディスタンス	ストレージ利用優先度
A-B	150	100	80	11
B-C	50	50	150	17
C-D	50	100	200	21
D-E	50	50	150	17
E-F	150	100	100	13
F-G	150	100	80	11
A-G	1000	200	10	11
B-E	600	300	200	24

(a)

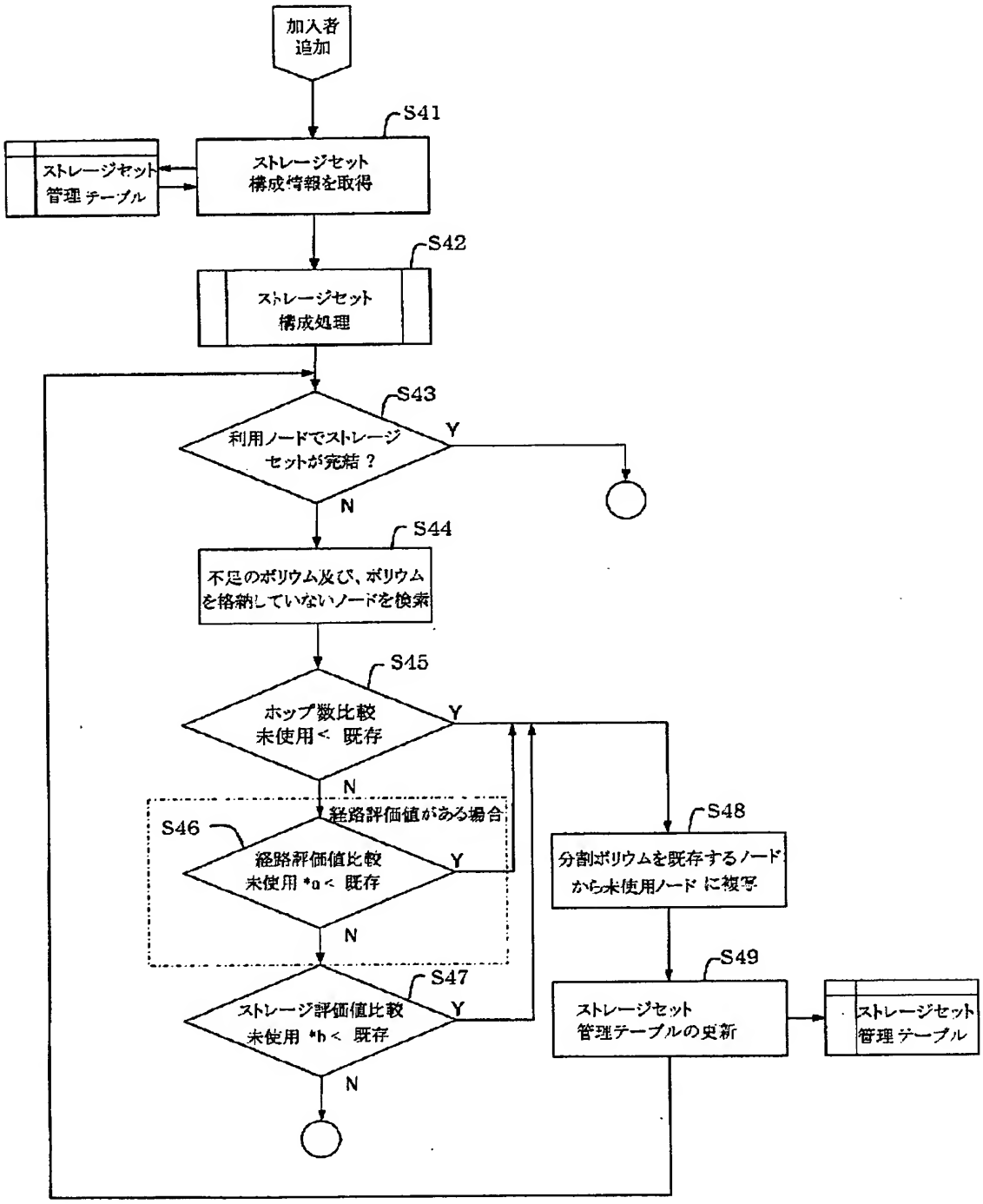
ストレージ評価テーブル

ノード	経路	式	ストレージ評価値
A			∞
B	A-B	= (A-B)	11.0
C	A-B-C	= {(A-B)+0.5(B-C)}/2	9.8
D	A-B-E-D	= {(A-B)+0.5(B-E)+0.33(D-E)}/3	9.6
E	A-B-E	= {(A-B)+0.5(B-E)}/2	11.5
F	A-G-F	= {(A-G)+0.5(F-G)}/2	8.3
G	A-G	= (A-G)	11.0

(b)

【図１５】

ノードへの利用者の追加処理を示すフローチャート



【図１６】

冗長化したデータの複製を格納する
ストレージの決定方法を説明する図

ノード	利用者A			利用者E		
	評価値	ホップ数	ボリウム	評価値	ホップ数	ボリウム
A	∞	0	a	17.8	2	←
B	11.0	1	b	24.0	1	←
C	9.8	2		16.3	2	
D	9.6	3		17.0	1	d'
E	11.5	2	c	∞	0	←
F	8.3	2		13.0	1	
G	11.0	1	d	10.8	2	←

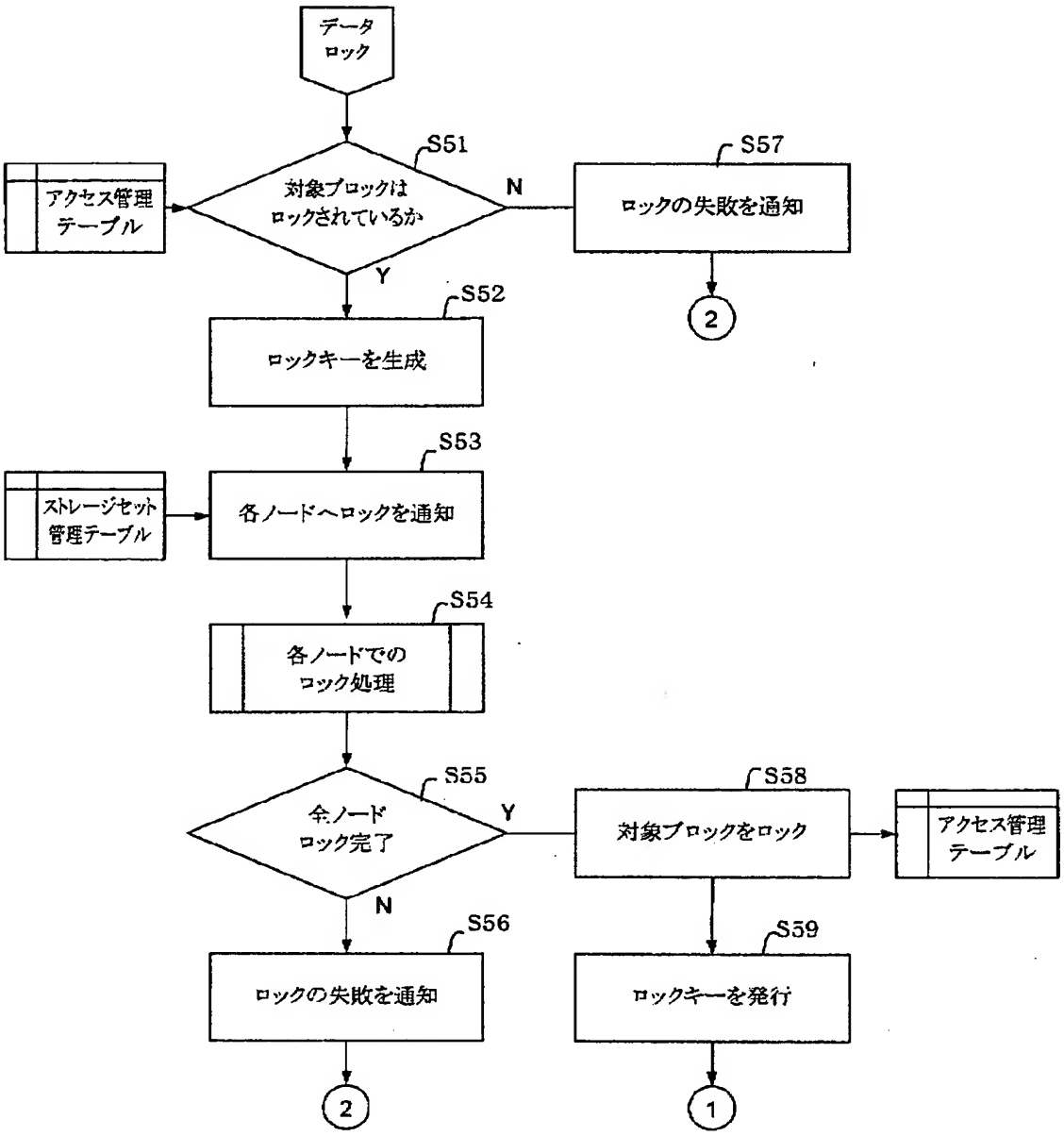
【図１７】

利用可能な複数のストレージから
最適なストレージセットを選択する処理を説明する図

ノード	利用者 A			利用者 E			利用者 C		
	評価値	ホップ数	ポリウム	評価値	ホップ数	ポリウム	評価値	ホップ数	ポリウム
A	∞	0	a	17.8	2	←	11.3	2	←
B	11.0	1	b	24.0	1	←	17.0	1	←
C	10.8	2		16.3	2		∞	0	a'
D	9.1	3		17.0	1	d'	21.0	1	←
E	11.5	2	c	∞	0	←	14.8	2	←
F	8.3	2	a'	13.0	1	←	10.9	3	←
G	11.0	1	d	10.8	2	←	8.4	3	←

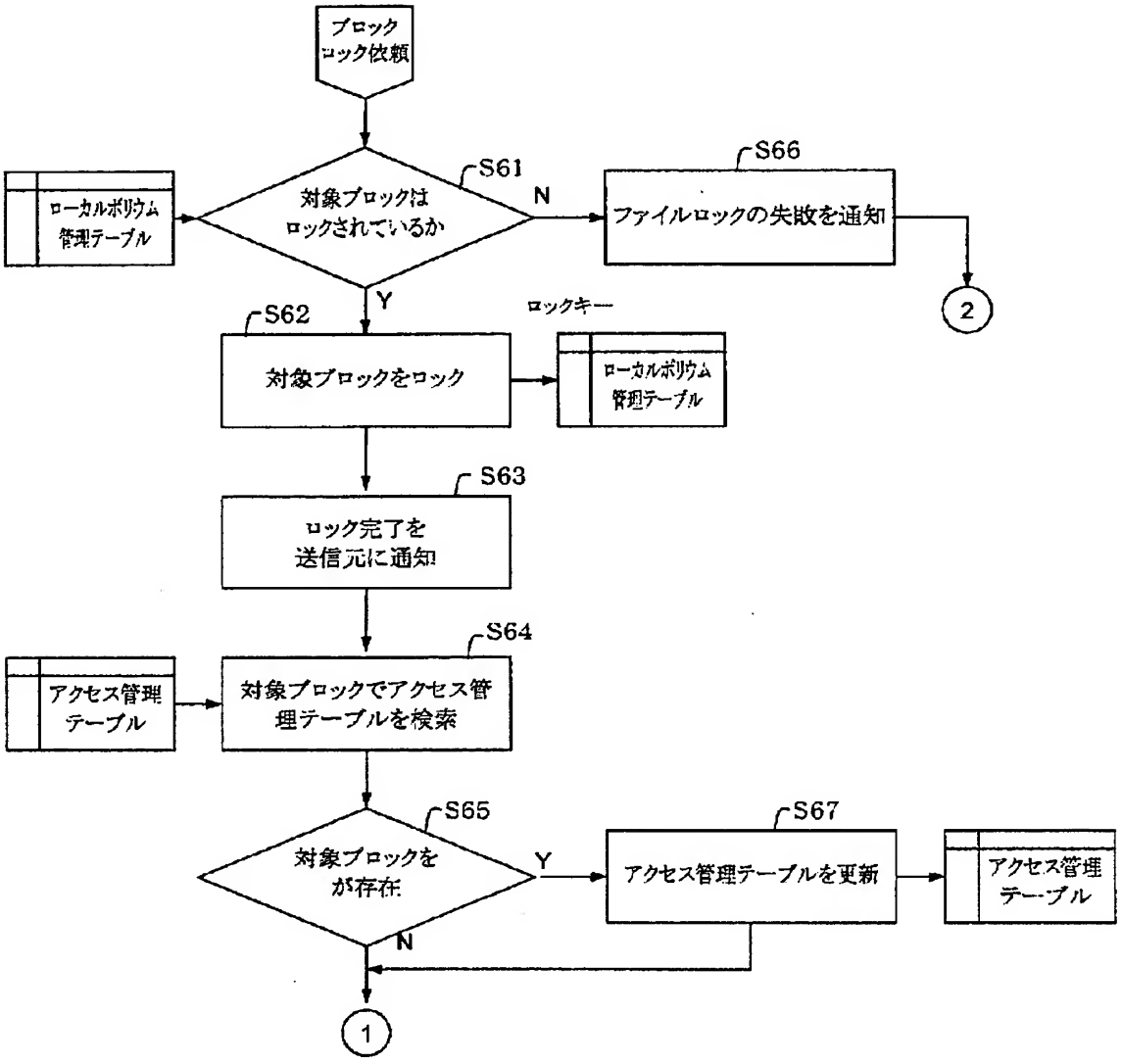
【図１８】

ロック処理を示すフローチャート（その１）



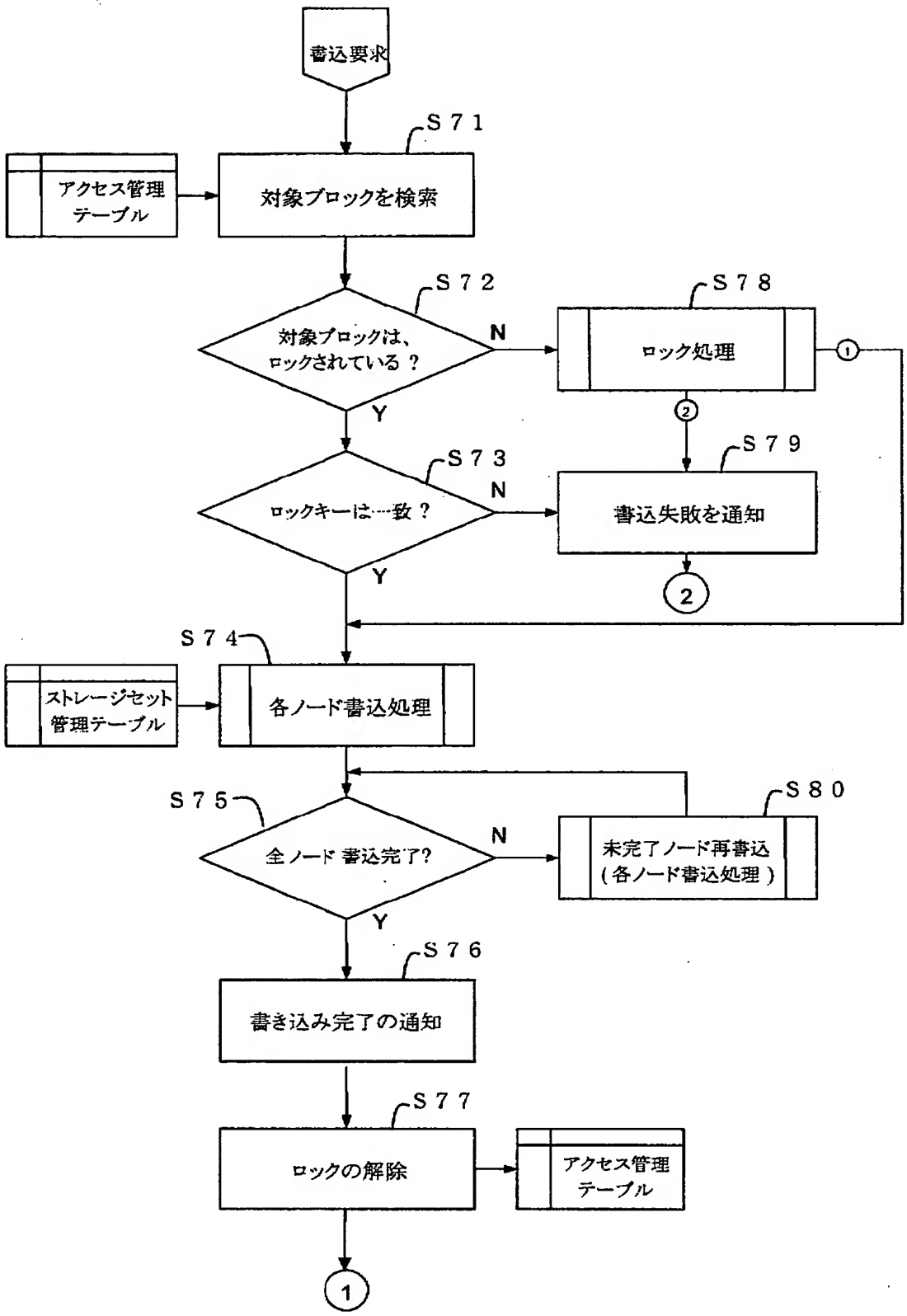
【図１９】

ロック処理を示すフローチャート(その2)



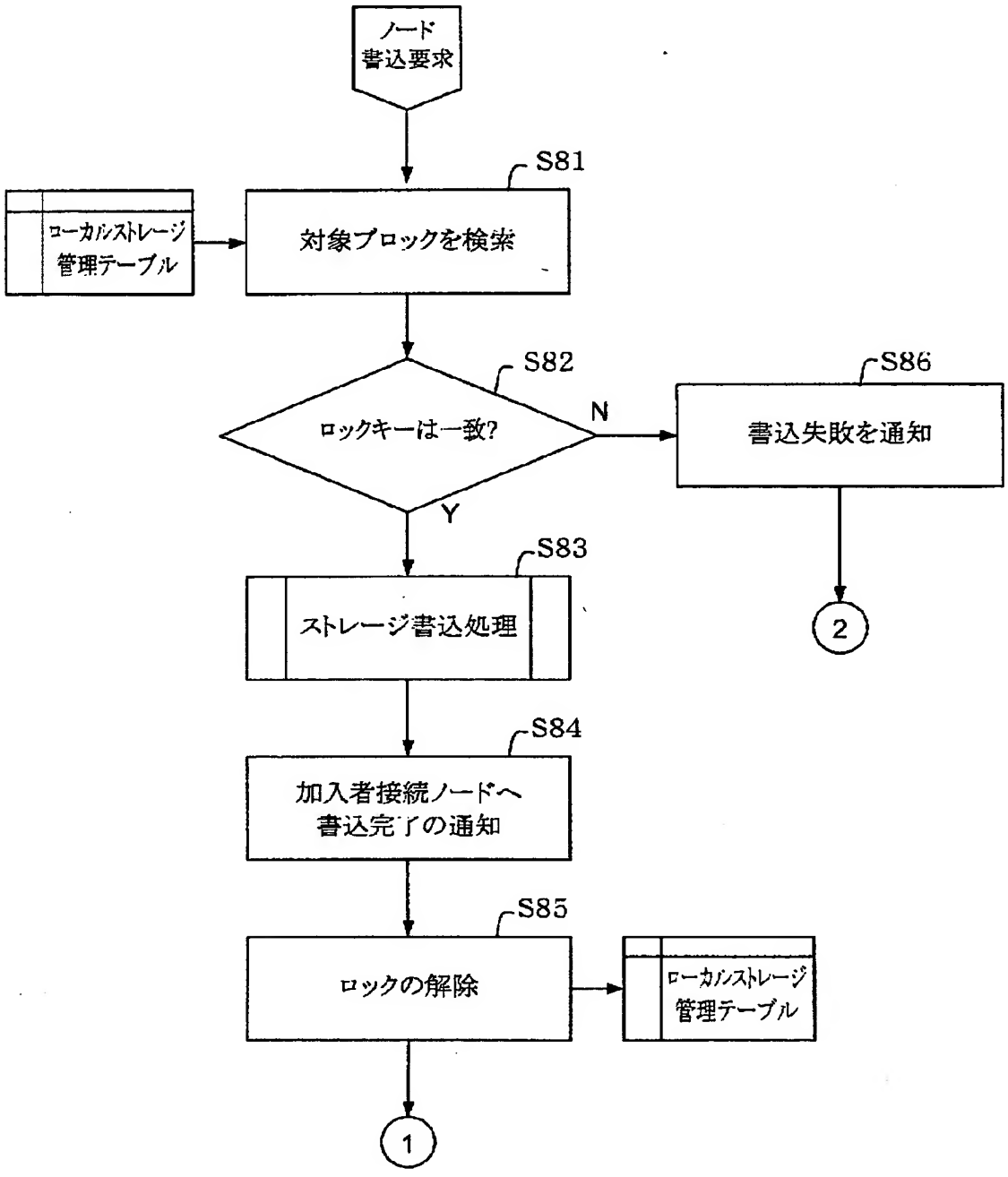
【図２０】

書き込み処理を示すフローチャート(その１)



【図２１】

書き込み処理を示すフローチャート(その2)



【図２２】

ストレージ内のデータ更新の際のロック処理を説明する図

ストレージセット番号	000010001	
ストレージセット アクセス番号	プロパティ	ロックキー
00000001	RWLO	0111344
00000002	RWLO	1124433
00000003	-W-O	
00000004	-W-O	
00000005	---O	
00000006	-W-O	

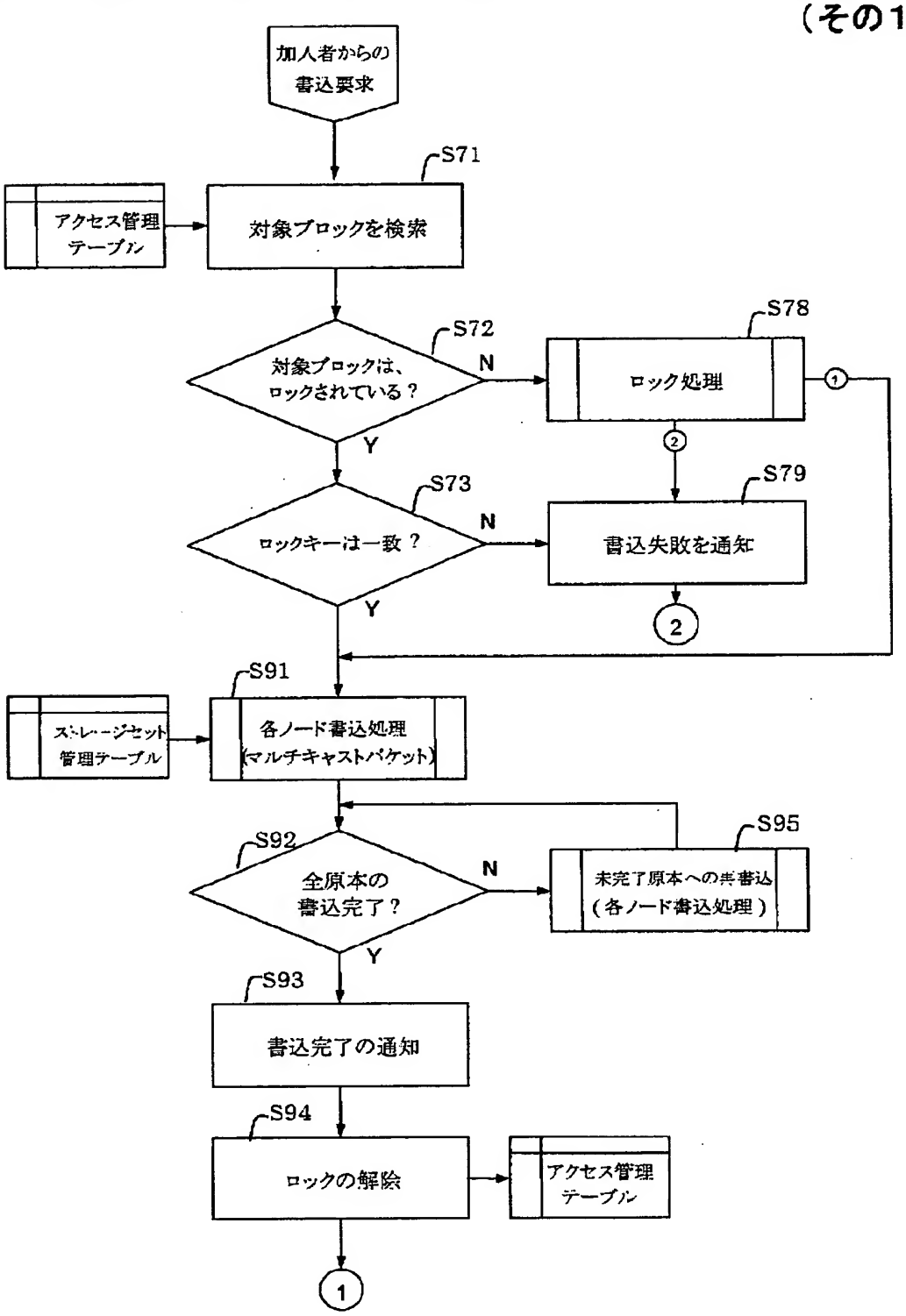
（ａ）

ローカルストレージ			ストレージセット管理情報	
ストレージ アクセス番号	プロパティ	ロックキー	ストレージセット 管理番号	ストレージセット アクセス番号
00000001	RWL		00010021	00040001
00000002	RW-	001223	02003001	00040004
00000003	RW-		00000000	00000000
00000004	R-F		01002111	01000100
00000005	RW-		00010021	00001562
00000006	RW-		00010021	00001563

（ｂ）

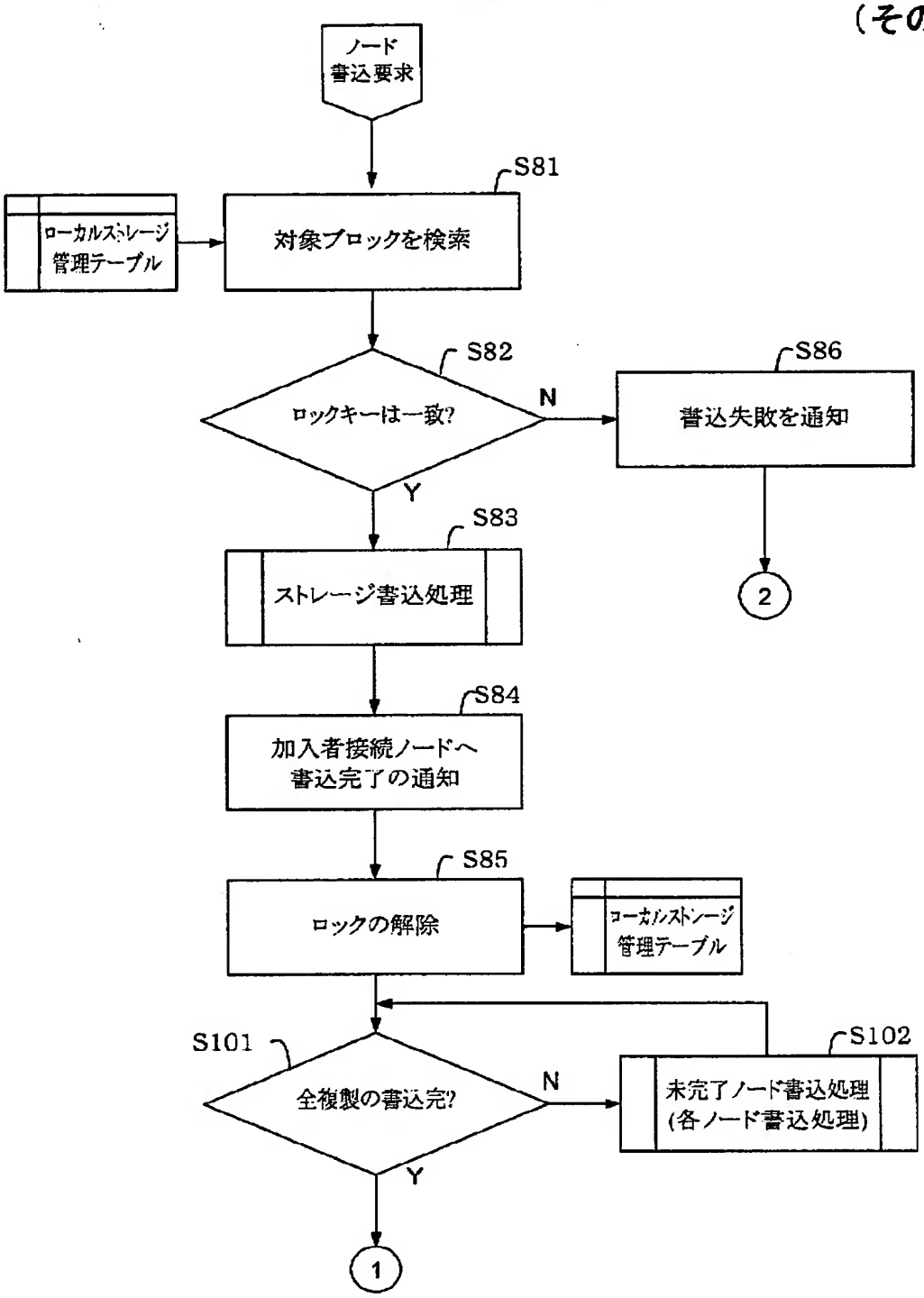
【図２３】

マルチキャストパケットを用いた書き込み処理を示すフローチャート（その1）



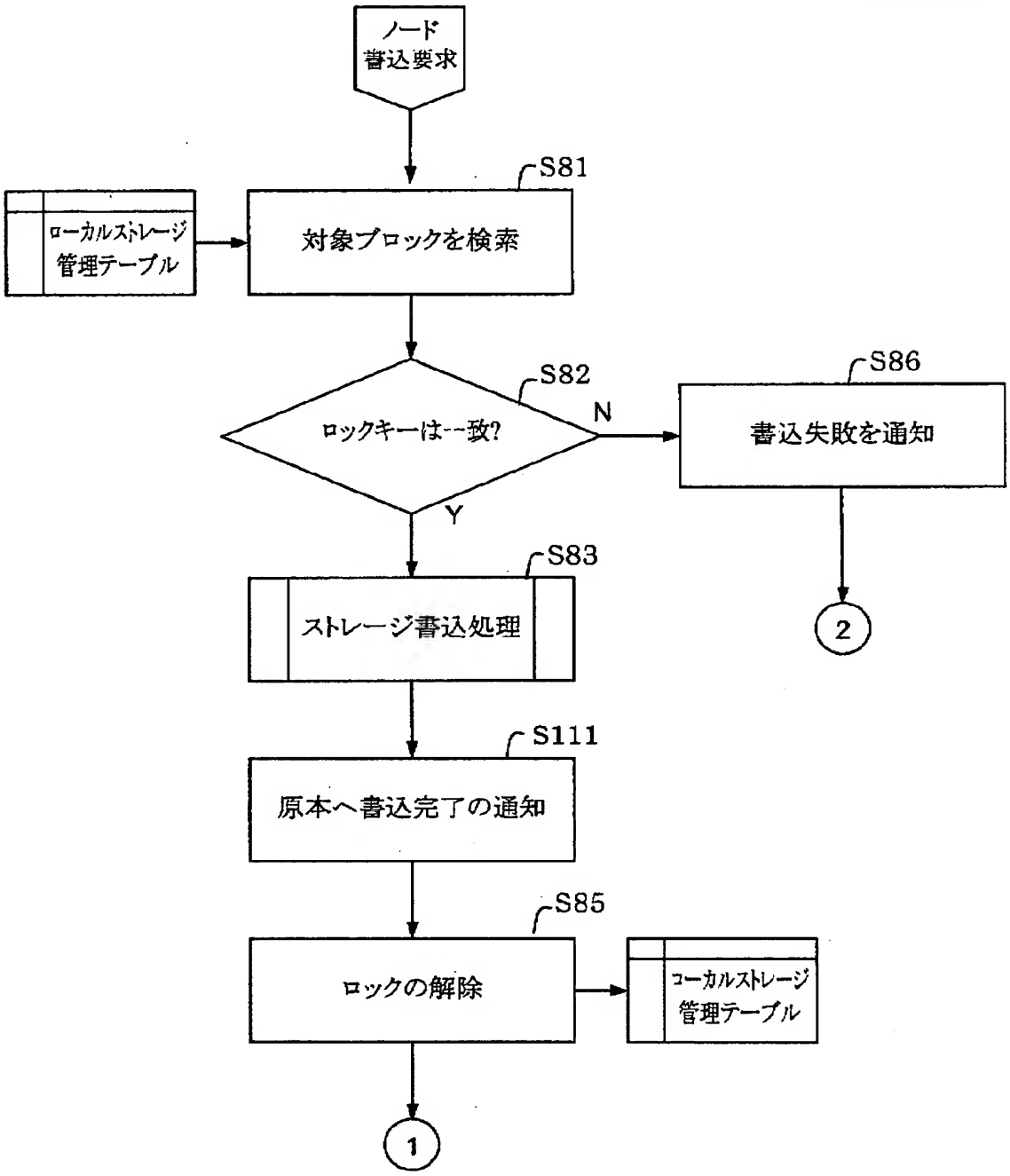
【図２４】

マルチキャストパケットを用いた書き込み処理を示すフローチャート（その2）



【図２５】

マルチキャストパケットを用いた書き込み処理を示すフローチャート（その3）



【図２６】

ボリウムの複製の作成方法を選択する処理を説明する図

ノード	利用者 A			利用者 E			利用者 C		
	評価値	ホップ数	ボリウム	評価値	ホップ数	ボリウム	評価値	ホップ数	ボリウム
A	∞	0	a	17.8	2	a	11.3	2	a
B	11.0	1	b	24.0	1	b	17.0	1	b
C	10.8	2	-	16.3	2	-	∞	0	-
D	9.1	3	d'	17.0	1	d'	21.0	1	d'
E	11.5	2	c	∞	0	c	14.8	2	c
F	8.3	2	a	13.0	1	a'	10.9	3	a'
G	11.0	1	d	10.8	2	d	8.4	3	d

【図２７】

一部のストレージに障害が発生した場合に、残りのストレージの中から最適なストレージを選択しなおす処理を説明する図

ノード	利用者 A			利用者 E			利用者 C		
	評価値	ホップ数	ボリウム	評価値	ホップ数	ボリウム	評価値	ホップ数	ボリウム
A	∞	0	a	17.8	2	a	11.3	2	a
B	11.0	1	b	24.0	1	b	17.0	1	b
C	10.8	2	a'	16.3	2	a'	∞	0	a'
D	9.1	3	d'	17.0	1	d'	21.0	1	d'
E	11.5	2	c	∞	0	c	14.8	2	c
F	8.3	2	a'	13.0	1	a'	10.9	3	a'
G	11.0	1	d	10.8	2	d	8.4	3	d

(a)

ノード	利用者 A			利用者 E			利用者 C		
	評価値	ホップ数	ボリウム	評価値	ホップ数	ボリウム	評価値	ホップ数	ボリウム
A	∞	0	a	17.8	2	a	11.3	2	a
B	11.0	1	b	24.0	1	b	17.0	1	b
C	10.8	2	a	16.3	2	a	∞	0	a
D	9.1	3	d'	17.0	1	d'	21.0	1	d'
E	11.5	2	c	∞	0	c	14.8	2	c
F	8.3	2	a'	13.0	1	a'	10.9	3	a'
G	11.0	1	d	10.8	2	d	8.4	3	d

(b)

【図２８】

ストレージが障害から復旧した際に
ボリウムの複製の作成方法を選択する処理を説明する図

ノード	利用者 A			利用者 E			利用者 C		
	評価値	ホップ数	ボリウム	評価値	ホップ数	ボリウム	評価値	ホップ数	ボリウム
A	∞	0	-	17.8	2	-	11.3	2	-
B	11.0	1	b	24.0	1	b	17.0	1	b
C	10.8	2	a	16.3	2	a	∞	0	a
D	9.1	3	d'	17.0	1	d'	21.0	1	d'
E	11.5	2	c	∞	0	c	14.8	2	c
F	8.3	2	a'	13.0	1	a'	10.9	3	a'
G	11.0	1	d	10.8	2	d	8.4	3	d

【図２９】

不要となった複製ボリウムを削除する処理を説明する図

ノード	利用者 A			利用者 E			利用者 C		
	評価値	ホップ数	ボリウム	評価値	ホップ数	ボリウム	評価値	ホップ数	ボリウム
A	∞	0	a	17.8	2	a	11.3	2	a
B	11.0	1	b	24.0	1	b	17.0	1	b
C	10.8	2	a	16.3	2	a	∞	0	a
D	9.1	3	d'	17.0	1	d'	21.0	1	d'
E	11.5	2	c	∞	0	c	14.8	2	c
F	8.3	2	a'	13.0	1	a'	10.9	3	a'
G	11.0	1	d	10.8	2	d	8.4	3	d

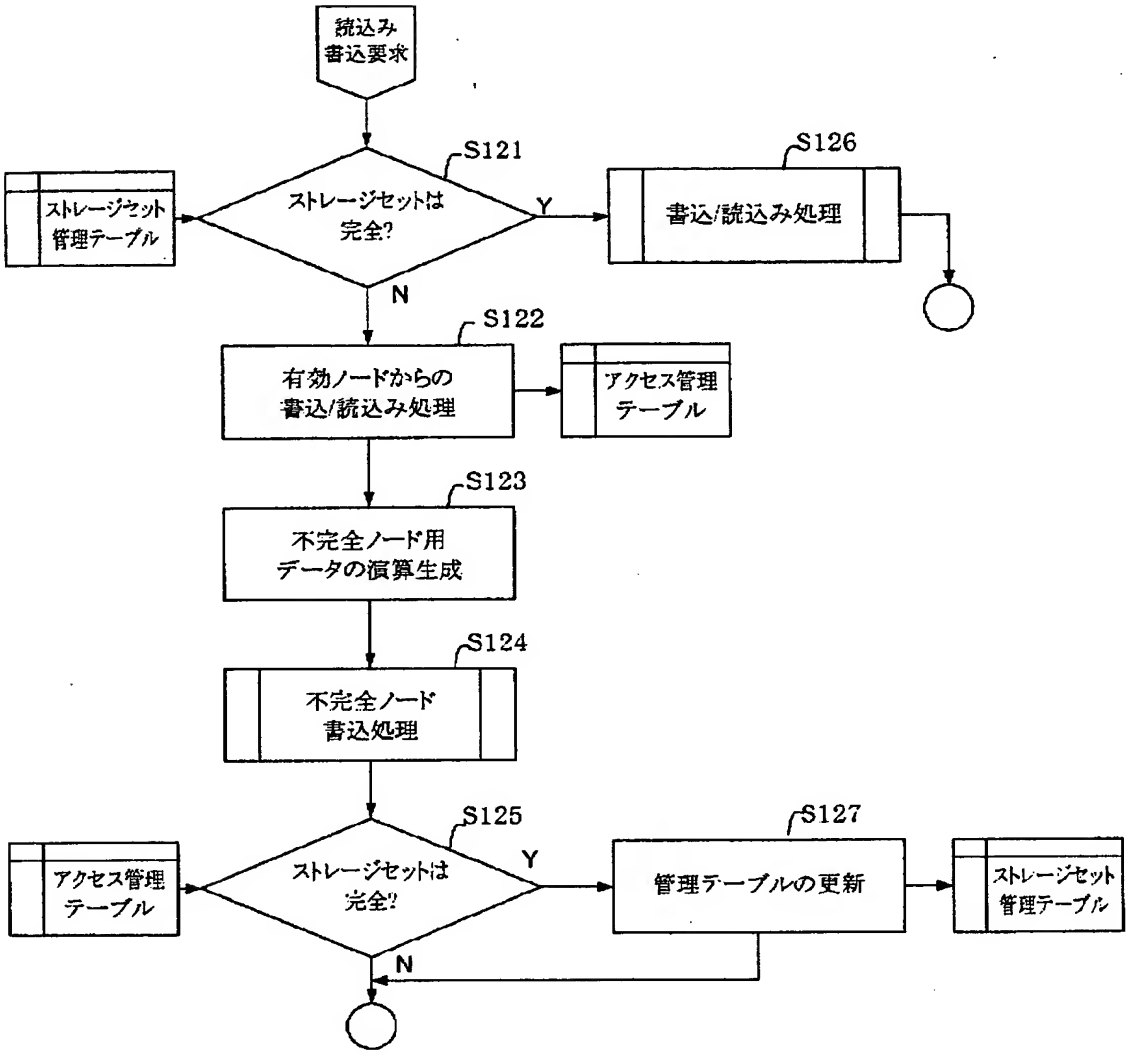
(a)

ノード	利用者 A			利用者 E			利用者 C		
	評価値	ホップ数	ボリウム	評価値	ホップ数	ボリウム	評価値	ホップ数	ボリウム
A	∞	0	a	17.8	2	a	11.3	2	a
B	11.0	1	b	24.0	1	b	17.0	1	b
C	10.8	2	a	16.3	2	a	∞	0	a
D	9.1	3	d'	17.0	1	d'	21.0	1	d'
E	11.5	2	c	∞	0	c	14.8	2	c
F	8.3	2		13.0	1		10.9	3	
G	11.0	1	d	10.8	2	d	8.4	3	d

(b)

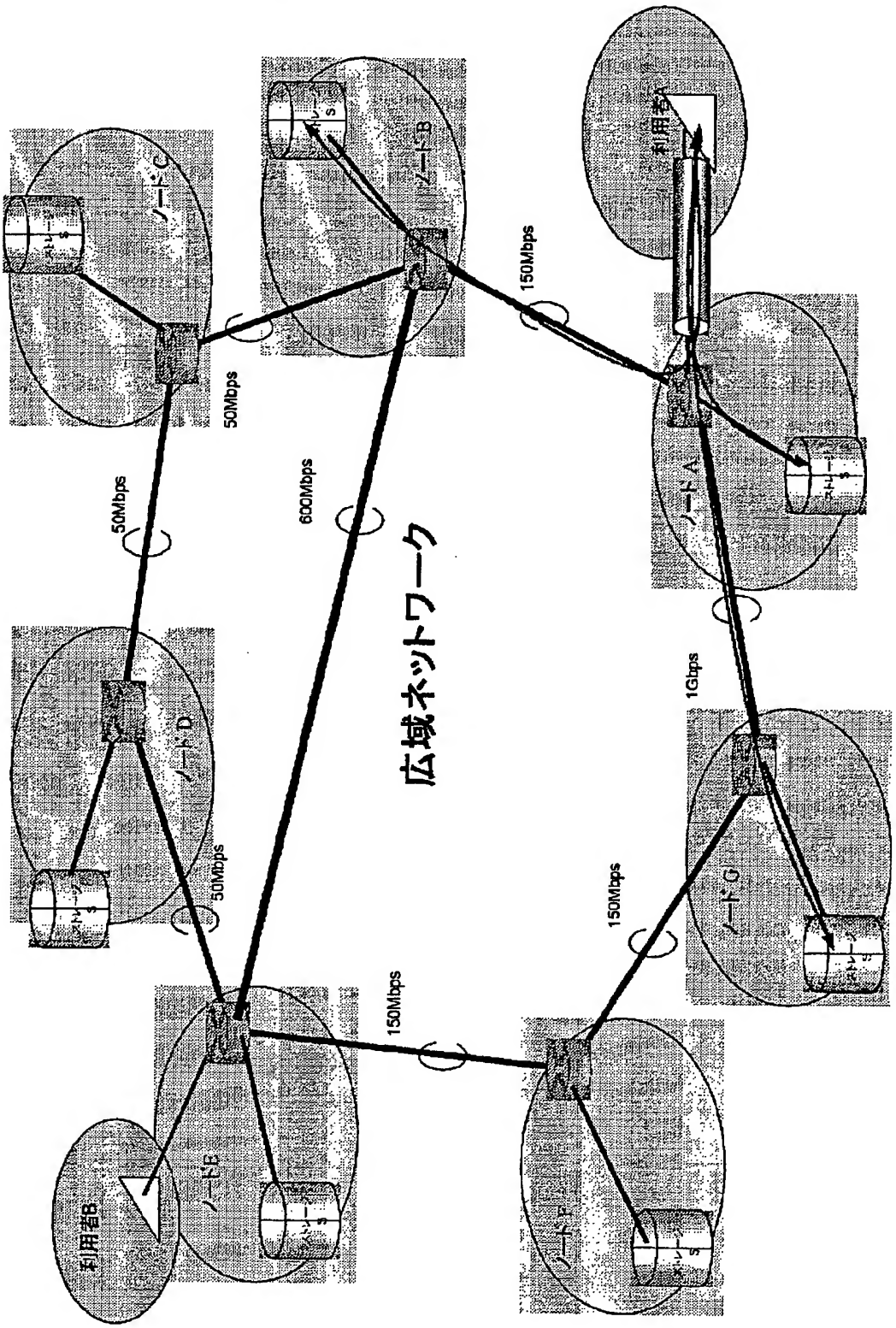
【図３０】

逐次にデータを書き込み又は再生する処理を示すフローチャート



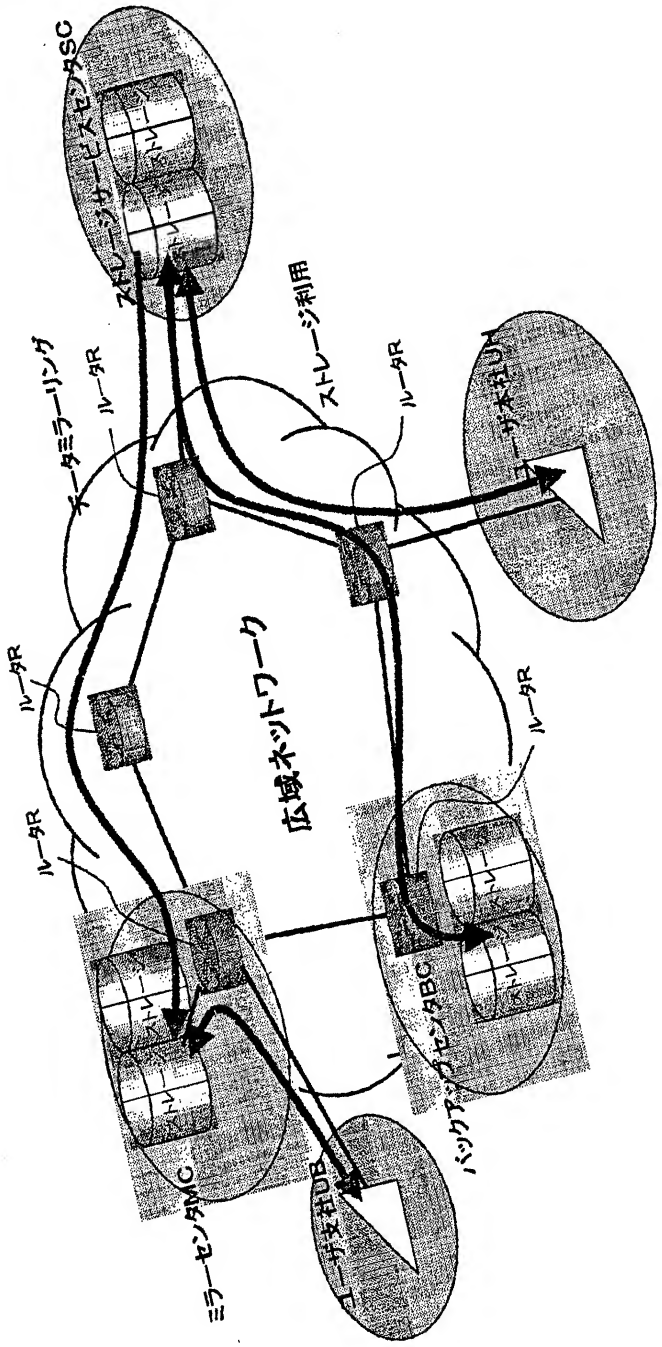
【図３２】

制御装置の機能を
利用者端末が備える場合を説明する図



【図３３】

従来の技術に係わる広域分散ストレージシステムの構成図



【書類名】　　要約書

【要約】

【課題】　データの冗長化に要するストレージ容量を低減しつつも、データのセキュリティを向上させ、且つ、回線を効率的に利用することが可能なＲＡＩＤを提供する。

【解決手段】　データを冗長化して複数のボリウムに分割し、各ボリウムを、ネットワークを介して分散配置された複数のストレージＳに分散して格納する制御を行う制御装置Ｃは、経路管理部５０４と、ストレージセット管理部５０５を備える。経路管理部５０４は、帯域幅、通信コスト及び書き込みを依頼するノードとストレージの間の物理的距離に基づいて、前記分散配置された各ストレージについて利用対象としての望ましさを示す評価値を算出する。ストレージセット管理部５０５は、前記評価値に基づいて前記分散配置されたストレージの中から複数のストレージを最適のストレージセットとして選択する。

【選択図】　　図３